# THE EPISTEMOLOGICAL IMPLICATIONS OF THE CAUSES OF MORAL BELIEFS

by

**Elizabeth O'Neill**

A.B., Biology, History, Brown University, 2008

Submitted to the Graduate Faculty of the

Kenneth P. Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

THE KENNETH P. DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Elizabeth O'Neill

It was defended on

May 5, 2015

and approved by

James Lennox, Professor, History and Philosophy of Science

Karl Schafer, Associate Professor, Philosophy

Kieran Setiya, Professor, Philosophy, Massachusetts Institute of Technology

James Woodward, Distinguished Professor, History and Philosophy of Science

Dissertation Advisor: Edouard Machery, Professor, History and Philosophy of Science

**THE EPISTEMOLOGICAL IMPLICATIONS OF THE CAUSES OF MORAL BELIEFS**

Elizabeth O'Neill, PhD

University of Pittsburgh, 2015

This dissertation investigates what the causes of moral beliefs indicate about the epistemic status of those beliefs. I argue that information about the causes of moral beliefs can tell us whether those beliefs track the truth, and that truth tracking is the primary epistemic property that should concern us in the moral domain. I formulate three novel debunking arguments that employ information about the causes of moral beliefs to support conclusions about truth tracking while minimizing normative assumptions. These arguments lead to the conclusion that harm-related moral beliefs that hinge on sympathy, moral beliefs influenced by disgust, certain political beliefs, and beliefs about punishment that are subject to the influence of extraneous emotions do not track moral truth. For each of these types of moral beliefs, information about the proximal causes of the moral belief supports epistemic conclusions. I compare the value of information about proximal and distal causes for assessing epistemic status: I argue that proximal causes are a superior source of information, but under certain conditions, we should take information about distal causes into account. In the case of beliefs about the fair distribution of resources, information about their proximal causes may be consistent with us tracking truth, but information about their distal, evolutionary origins tell us that we lack reason to think that we track their

truth. Thus, using empirical information about the causes of moral beliefs, I offer selective debunking arguments for five types of moral beliefs.

## TABLE OF CONTENTS

**PREFACE** ..... **IX**

**1.0 INTRODUCTION** ..... **1**

**1.1 My debunking arguments in relation to the types of debunking arguments in the literature** ..... **4**

**2.0 TRUTH TRACKING AS THE MOST IMPORTANT EPISTEMIC PROPERTY IN THE MORAL DOMAIN** ..... **12**

**2.1 Roush's account of truth tracking** ..... **14**

**2.2 The practical value of true belief, absence of false belief, and truth tracking** ..... **17**

**2.3 The relationship between matching, truth tracking, and safety, and how to investigate the probability of matching** ..... **22**

**2.4 The practical value of matching and truth tracking in the moral context** ..... **24**

**2.5 Preliminary Worries** ..... **30**

**2.6 Objections** ..... **33**

**2.7 Other epistemic properties** ..... **36**

**2.8 What determines sensitivity, adherence, and the probability of matching? The relationship between causes and truth tracking** ..... **41**

**3.0 THE PRIORITY OF PROXIMAL CAUSES AND THREE PROXIMAL DEBUNKING ARGUMENTS** ..... **44**

**3.1 Proximal causes as a superior source of evidence about epistemic status** ..... **44**

**3.2 The Priority of Proximal Causes ..... 46**

**3.3 Three debunking arguments ..... 52**

**3.3.1 The Malfunctioning Component Argument: Sympathy ..... 52**

**3.3.2 The Argument from Inconsistent Variability ..... 60**

**3.3.3 The Epistemic Significance of an Inconsistency in Levels of Variability ..... 62**

**3.3.4 How We Can Identify Inconsistent Levels of Variability ..... 67**

**3.3.5 The Inconsistent Levels of Variability Argument Applied in the Moral Domain ..... 70**

**3.3.5.1 Overview ..... 70**

**3.3.5.2 Attitude Hypervariability: Case 1: Variability in attitudes on moral questions in politics over a lifetime ..... 71**

**3.3.5.3 Attitude Hypervariability: Case 2: Variability in attitudes about punishment influenced by externally induced emotions ..... 75**

**3.3.5.4 Attitude Hypovariability: The Incorrigibility of Disgust ..... 80**

**3.3.6 The Inappropriately Improbable Belief Argument ..... 89**

**3.3.7 Equiprobable beliefs ..... 91**

**3.3.8 Highly improbable belief ..... 91**

**3.4 Objections ..... 97**

**3.5 The scope of debunking arguments ..... 98**

**3.6 Conclusion ..... 100**

**4.0 THE DISTAL CAUSES OF FAIRNESS NORMS ..... 102**

**4.1 Introduction ..... 102**

**4.2 The inappropriately improbable process argument ..... 104**

**4.3 Fairness norms ..... 106**

**4.4 Development of Western fairness intuitions ..... 111**

**4.5 Evolution of fairness intuitions ..... 114**

**4.6 Objections ..... 120**

**4.7 Conclusion ..... 121**

**5.0 CONCLUSION ..... 122**

**BIBLIOGRAPHY ..... 124**

## PREFACE

I am profoundly grateful for the support and guidance of my advisor, Edouard Machery. I am also indebted to the members of my committee, James Lennox, Karl Schafer, Kieran Setiya, and James Woodward, for their generous feedback and advice.

I would like to extend particular thanks to the members of my research support group over the last five years, Kathryn Tabb, Julia Bursten, Aleta Quinn, Marcus Adams, Bihui Li, Catherine Stinson, Karen Zwier, Peter Distelzweig, and Jason Byron. Many thanks also to Yoichi Ishida, Taku Iwatsuki, and Joseph McCaffrey. I am grateful to Zina Ward, Jasmin Özel, Annalisa Paese, Jennifer Zamzow, Adam Marushak, Robert Steel, Joshua Hancox, David Colaço, Katie Elliott, Daniel Kelly, Victor Kumar, Brandon Williams, and Daniel Storey for helpful feedback on past versions of portions of this dissertation. I am also grateful to Bob Barnard, Liam Kofi Bright, Jake H. Davis, Gregory Gandenberger, Aaron Glieberman, Kareem Khalifa, Micah Krafcik, Stephen Makin, Michael Miller, Aaron Novick, Sean O'Neill, Patricia Rich, Lauren Ross, Sherrilyn Roush, and many others for helpful conversations on topics related to the dissertation.

I presented parts of this project at two Graduate Student Work in Progress talks at the University of Pittsburgh, the 2014 meeting of the Philosophy of Science Association, the 2014 meeting of the Conference on Value Inquiry: Evolution and the Foundations of Ethics, the 2014 meeting of the Midsouth Philosophy Conference, the 2014 meeting of the Southern Society for

Philosophy and Psychology, and the 2013 meeting of the International Society for the History, Philosophy, and Social Studies of Biology, and I am grateful to audiences at each of these talks for productive discussions.

I would like to thank Joann McIntyre and Natalie Schweninger for their encouragement and expert administrative assistance.

Finally, I want to express my appreciation for a fellowship from the Dolores Zohrab Liebmann Fund that supported my research in 2014-2015, for conference travel support from the Wesley C. Salmon Fund, and for access to the University of Pittsburgh Adolf Grünbaum Philosophy Reading Room.

# 1.0 INTRODUCTION

This dissertation considers the significance of the causes of moral beliefs for the epistemic status of those beliefs. Recent empirical research offers us evidence about the variety of psychological and biological factors that influence our beliefs about moral principles and the judgments we make about moral situations. Such research raises questions about whether the influence of certain types of causes should lead us to reduce confidence in our moral beliefs.

In the second chapter, I lay the epistemological groundwork for the rest of dissertation. I argue that in the moral domain, we should primarily be concerned with fulfilling any obligations we might have. For this purpose, our primary epistemic concern should be whether we are likely to have the right beliefs about claims related to our obligations—such as that we ought to relieve the suffering of others if it comes at little cost to us, or that we ought to give up our seat on the bus under certain circumstances. The tendency to have the right beliefs about these claims can be captured in terms of truth tracking, an epistemic property that has been most helpfully elaborated by Sherrilyn Roush: roughly, a person tracks the truth if she has a high probability that she believes a claim when it is true and a high probability that she disbelieves the claim when it is false. I defend the view that truth tracking, regardless of its relation to knowledge and justification, is the primary epistemic property that should concern us in the moral domain. In making this point, I follow Alston's practice of focusing on desirable properties of beliefs rather than knowledge or justification (2005). I argue that with information about the causes of beliefs,

we can evaluate whether we track the truth of the propositions we believe.

In the third chapter, I argue that more proximal causal processes like emotions and reasoning processes are a better source of evidence for assessing whether we track the truth than distal process like cultural or evolutionary histories. As a result, investigations into the epistemic status of our beliefs using information about their causes should begin with the information we have about proximal psychological faculties. In this chapter I also develop three distinct argumentative strategies for debunking. The first strategy is to identify a step in the belief production process that we can show empirically is not performing the function it would need to perform for the process to track the truth, and to infer from this that our belief forming process does not allow us to track the truth. I call this the malfunctioning component argument. The second strategy, using metaethical rather than normative assumptions, is to show that the variability in the truth value of a proposition is inconsistent with the variability in our attitudes about the proposition. I call this the argument from inconsistent variability. Third, I characterize a debunking strategy based on the improbability of having the belief one happens to have.

I present several cases to illustrate these three arguments as well as to illustrate the point that we can draw skeptical conclusions on the basis of information about proximal causes alone. My first case, to illustrate the malfunctioning component argument, examines the influence of sympathy on moral beliefs. I begin by supposing that if sympathy were to contribute to truth tracking moral judgments, it would do so by picking out entities that are suffering. I appeal to empirical work to argue that we cannot rely on sympathy to perform this function because it is highly vulnerable both to false positives—cases where we think there is suffering but there is not, which may occur in response to robots, for instance—and false negatives—cases where we think there is no suffering but there is, which may occur as a result of out-group biases, for

instance. Second, I argue that certain properties of disgust mean that moral beliefs influenced by disgust are not tracking the truth of the propositions in question. Empirical work shows that disgustingness is "sticky"—it is too transmittable, too easily associated with diverse actions and objects, and it is too incorrigible, too resistant to revision once it is associated with an action or object. The incorrigibility of disgustingness illustrates one version of the argument from inconsistent variability—attitude hypovariability, in which one's attitudes are too invariable in comparison with the variability we expect in the truth value of the proposition. The transmissibility of disgustingness illustrates the argument from inappropriately improbable belief. In my third and fourth cases, I illustrate the other version of the argument from inconsistent variability—attitude hypervariability, in which one's attitudes vary too much in comparison with the variability we expect in the truth value of the proposition—with the case of moral political beliefs that are likely to change over one's lifetime and with the case of judgments about punishment that are likely to vary as a result of variation in extraneous emotions. In each of the cases I examine in this chapter, we are able to infer that our beliefs are not truth tracking on the basis of just an examination of the influence of proximal causes; it is unnecessary to also look at more distal causes in these cases.

In the fourth chapter, I consider a case where even if information about proximal and developmental causes of beliefs is consistent with truth tracking, information about distal causes should lead us to reduce confidence in our beliefs. This is the case of beliefs about fairness, specifically about resource distribution. For instance, one might think that food should be distributed according to need, or that it should be distributed according to who contributed the most to obtain it. This case illustrates a higher order version of the inappropriately improbable belief argument, the third debunking argument that I discussed in the second chapter, which here

becomes the inappropriately improbable *process* argument. Employing a variety of empirical evidence, I argue that the evolutionary origins of intuitions about fair distribution of resources show that it was improbable that the particular fairness faculty humans currently have would evolve, and the intuitions humans have about fair distributions of resources could easily have been different. In combination with metaethical assumptions from a realist or commonsense view of the nature of morality, the improbability of people's beliefs in this domain puts us in a poor epistemic situation, and we should reduce confidence in these beliefs.

Even if we start with an assumption that we have a prima facie reason to trust our moral faculties or that moral epistemology must be biased toward the truth to avoid skepticism, each of these debunking arguments lead to the conclusion that we are in a poor epistemic position.

## 1.1 MY DEBUNKING ARGUMENTS IN RELATION TO THE TYPES OF DEBUNKING ARGUMENTS IN THE LITERATURE

There are a variety of views on how causes bear on the epistemic status of beliefs (e.g. White 2010, Cohen 2000, Bedke 2009, 2014, Goldman 2014, Vavova ms.). On one side, some argue that causes do not affect the epistemic status of beliefs and/or do not reveal anything about their epistemic status. On the other side, there is an assortment of causal debunking arguments.[1] In

[1] For additional discussions of causal debunking arguments, see Griffiths and Wilkins n.d., Ruse 1986, Shafer-Landau 2012, Sher 2001, Tersman 2008, Machery and Mallon 2010, Brosnan 2011, Kahane 2011, 2013, Clark-Doane 2012, Schafer 2013, Schechter 2013, Setiya 2013, Berker

this dissertation, the debunking arguments I introduce differ from other characterizations of genealogical debunking arguments in the literature.

Perhaps the most familiar causal debunking argument is what has been called the best explanation argument, in which one shows that the best explanation for a belief nowhere appeals to the truth of the believed proposition. Varieties of this argument have appeared in e.g. Harman 1977, 1986, Ruse 1986, Street 2006, and Joyce 2006. For instance, one argument says that if an explanation (or a complete explanation or our best explanation) of a belief need not involve the fact of the matter, and if we adhere to a principle of parsimony, we (or just realists) lack reason to retain the belief. In this dissertation, I will not address debunking arguments that arise from explanations of beliefs or from our difficulties explaining moral knowledge or the purported reliability of our moral beliefs.

A second form of etiological debunking argument shows that a belief is the product of a process known to be unreliable. Nichols characterizes this sort of argument in his (2014); he proposes that there are two types of etiological debunking arguments, one of which is the explanation approach that I described above and the other is the process debunking argument. Nichols characterizes the process debunking argument as an argument in which one concludes that a belief is unjustified by showing that a belief is the product of a process that is known to be epistemically defective (e.g., distorting or unreliable), such as dreaming or wishful thinking

---

2009, Street 2006, 2008, Copp 2009, Wielenberg 2010, Enoch 2010, Skarsaune 2011, Fraser 2014, and Vavova 2014.

(2014, 2; see also Sinnott-Armstrong 2008 for discussion of this sort of argument).[2] Another account of causal debunking arguments that involves a premise about the epistemic defectiveness of one's process is Kahane's (2011) reconstruction of evolutionary debunking arguments, which requires the premise that one's process is off-track (106). To advance these sorts of arguments, one must assume or supply reasons to think that one's moral belief process is epistemically defective. This is harder to do in the moral domain than in other domains. One way to support a premise about the unreliability or truth tracking failure of a process is to employ the irrelevant factors argument, and show that a key component of the process that produces the belief is a morally irrelevant factor.

The irrelevant factors argument, though related to the process debunking argument, is distinct, and so constitutes a third type of debunking argument. This argument supplies an account of some of the causal influences on a belief and shows that some of the important influences on the formation of the belief are not responsive to the fact of the matter and does not contribute to the process's responsiveness to the fact of the matter as a background condition that

---

[2] Schafer has also made a related point, that one could conclude one should distrust one's belief if one knows that it is due to an unreliable process *or* if one knows that for the belief to have a high probability of being true, the belief must be the product of one of a small set of processes that one knows to be reliable, and that the belief is not produced by one of those processes. (Perhaps this implies that one knows that the other processes are unreliable, but the argument is driven by the idea that the truth in a particular domain is tricky to get at, and there is only a small set of certain sorts of processes that can be relied upon, and so the argument seems potentially different in a way that might involve fewer normative assumptions.)

helps other factors be responsive to the fact of the matter. The irrelevant factors argument has been presented in a number of different ways. Vavova (2014), for instance, says the influence of irrelevant factors is bad when it is distorting. Similarly, Mallon and Doris (2013) discuss the problem produced when automatic processes are either sensitive to factors unrelated to morality or insensitive to factors that are relevant to morality. (They also mention in passing the possibility of a very different debunking argument, which operates by showing that moral responses might not have the right type of stability needed for them to be reliable. This is the sort of concern that my argument from inconsistent variability explicates.) In another domain, Swain et al. (2008) discuss the epistemic significance of epistemic intuitions tracking factors that are irrelevant to the question at hand (140-141). More precisely, we might interpret the irrelevant factors argument this way: (1) S believes *p* as the result of a process G, which contained factor A; (2) S was unlikely to believe *p* given the absence of factor A in process G, and S was likely to believe *p* given the presence of factor A in process G; (3) factor A is not an indicator or determinant of the truth value of *p* and factor A does not serve as a background condition that enables other component factors of the process to function as indicators of A. From these premises, one can conclude that process G is an unreliable source for obtaining attitudes about *p* and that S is not tracking the truth of *p*.

In this argument, though, one has to show that the allegedly irrelevant factor is in fact not an indicator of the truth value of p and is not a background condition conducive to other component factors tracking *p*. This usually involves appealing to normative assumptions. For instance, when Greene (2008) writes that deontological intuitions "reflect the influence of morally irrelevant factors and are therefore unlikely to track the moral truth," he has to claim that

particular factors are morally irrelevant (70).[3] Our deontological intuitions, Greene argues, are the product of an automatic process that produces different judgments depending on variation in factors such as personal force and spatial distance, which Greene claims are not morally relevant.[4] Consequently, we must decrease our confidence in moral beliefs formed by these automatic processes. However, the assumption that personal force and spatial distance is not a determinant of whether an action is right or wrong is not obvious and has attracted objections (Berker 2009, Kahane 2013).

Setting the irrelevant factors argument aside, an alternative way to show that a moral process is epistemically defective is to show that it involves a component process or a feature that is epistemically defective across contexts. However, as Nichols notes, "sometimes processes are defective in some contexts and not others" (2014, 7). Thus, it is difficult, particularly in the moral domain, to support the claim that one's belief processes are epistemically defective. There are also some arguments that might be counted as weaker versions of the irrelevant factors or process debunking arguments: these arguments just show that the belief of interest is the product

[3] Singer (2005) advances a similar sort of argument.

[4] Greene might be supporting the claim that these factors are morally irrelevant with his point that our sensitivity to those factors is a product of evolution, which does not track the truth of moral propositions. If so, he is relying on a different, disputed assumption about evolution's failure to track moral facts, or must supply a different argument to support the claim. (His 2014 argument for the unreliability of automatic processes as a result of their evolutionary history offers a possibility).

of a process that we lack reason to think is tracking the truth (one could interpret some Street- and Joyce-type arguments this way).

To the extent that the irrelevant factors argument and the process debunking argument rely on normative assumptions—for instance, about whether given factors are morally relevant—they appear to be varieties of what Shafer-Landau (2013) calls the "knowledge-based genealogical critique." Bedke (2014), too, identifies a class of debunking arguments like this. He distinguishes between two types of debunking arguments—those that proceed by establishing that our substantive moral beliefs are true by some sort of problematic coincidence, and those that undermine the status of subsets of beliefs on the grounds that they are influenced by "nefarious forces," a judgment that is made by appealing to "privileged moral beliefs."

My arguments from inconsistent variability and inappropriately improbable belief rely on neither a premise about unreliability nor a premise about the lack of correlation between factor A and *p*. My argument might fall into a category that Shafer-Landau calls the "agnostic genealogical critique," which seeks to show "without any assumption of where truth in a given domain lies," that beliefs that result from some process are unlikely to be true (2012, 2). However, the arguments he discusses in this category rely on an assumption that there is no connection between the fact and the causal factor influencing beliefs. My argument avoids even this sort of assumption. One might think that the dysfunctional component argument that I advance falls into the category of knowledge-based genealogical critique, but in fact that argument uses a normative assumption only with the aim of identifying an inconsistency in moral claims—a tension between the idea that a certain set of moral beliefs are true and responsive to something we take to be morally relevant in the world, and the fact that the mechanism in our process that would have to track that property for the process to be truth

tracking is dysfunctional. Thus, the debunking arguments I formulate in this dissertation minimize assumptions about privileged moral beliefs and assumptions about which actions are right or wrong.

This project also relates to what Pritchard (2005) has characterized as the "epistemological luck research program." In the literature on knowledge and justification, many people have talked about the conditions under which luckiness prevents one's beliefs from being justified or constituting knowledge. The question of luckiness is linked to the improbability of one's beliefs. Not all situations in which one's belief is improbable deprive one of knowledge. For instance, philosophers tend to think that if one obtains a rational faculty or some other reliable process improbably or only by luck, this is not an epistemic threat—e.g. no threat to one's belief constituting knowledge or being justified. (See e.g. White's 2010 comparison of coin flips in the head that determine belief and coin flips outside the head that determine whether one obtains a rational process for acquiring belief.) However, in my third chapter, I argue that under certain conditions, it being improbable that one would obtain a process that produces the beliefs one has on a topic undermines one's reason for thinking that one tracks the truth.

White observes that the epistemic problem posed by the causes of one's beliefs is sometimes presented as a problem of luck, contingency, chanciness, and precariousness of one's belief, and other times presented as a concern about inevitability and predictability of one's belief. He says, "if both factors by themselves raise an epistemological problem then it would seem to be a matter of damned if you do and damned if you don't" (579). Looking at a particular belief or set of beliefs, he is surely correct; we cannot say that the belief was both too probable and not probable enough. And if we were evaluating the epistemic status of all of our moral

beliefs together, only one of these features at a time could be a problem.[5] If we are examining different moral propositions or classes of moral propositions, and we have different expectations about the probabilities of those different beliefs, or if we have different background assumption about how the truth value of those moral propositions varies over time, we can be worried about the luckiness of some beliefs and the inexorability of others, and vulnerability to inappropriate loss of some beliefs and inappropriate incorrigibility of others. Each of the debunking arguments I advance is a selective debunking argument that may be applied to just a subset of moral beliefs, and etiology will produce one set of epistemological problems for some of these moral beliefs and a different set for the others.

[5] White explains the pull of these conflicting problems as due to the chanciness of one's belief occurring earlier in the causal process and the fixity of one's belief occurring once various unlikely causal factors are in place. He says, "Of course there is no inconsistency between these two factors, they just occur at different stages. It is a contingent matter which causes are in place that make it inevitable that you will believe this or that" (579). I do not think this is the right analysis of naïve worries about etiology nor the solution to those worries, but this does not matter for my point above.

## 2.0 TRUTH TRACKING AS THE MOST IMPORTANT EPISTEMIC PROPERTY IN THE MORAL DOMAIN

In this dissertation I investigate what information about the causes of moral beliefs reveals about the epistemic status of those beliefs. There are several epistemic properties that might interest us: we could investigate whether moral beliefs constitute knowledge, whether they are justified or warranted, "safe," the product of reliable processes, or whether they track the truth. Indeed, there are a variety of properties we might want our beliefs to have. I follow William Alston (2005) in thinking that in different contexts, these different epistemic desiderata may be more or less valuable (176). In this chapter I will argue that *tracking the truth* of certain propositions is the epistemic property that matters most for completion of any moral obligations that we may have, and, as a result, when we are investigating the epistemic status of beliefs in the moral domain, truth tracking is the most important property to examine. Further, I will argue that the degree to which we track the truth of a proposition is affected by the causes that influence whether we believe or do not believe the proposition and by the connection between those causes and the fact of the matter or the truth makers of the proposition. Consequently, to find out whether we track the truth, we should look into the causes that influence our beliefs.

Much of the moral debunking literature has been presented in terms of truth tracking—Sharon Street presented her Darwinian dilemma for the realist in terms of whether evolution is a

process that track truth and whether our moral beliefs track the truth.[6] Several respondents to Street, such as Kahane, Enoch, and Brosnan, and others who have advanced evolutionary debunking arguments, such as Griffiths and Wilkins, have also used the language of truth tracking.[7] Street says that she borrowed the language of truth-tracking from Nozick, but she and others tend to use the idea somewhat loosely, and do not invoke the specifics of Nozick's account (Street 2006, 159 n.26). People have offered a number of characterizations of the epistemic problem posed by the influence of certain causes on moral beliefs. In this dissertation, I employ Roush's account of truth tracking in terms of conditional probabilities. In the next chapter, I show how information about the causes of moral beliefs can be employed in three types of debunking arguments. For each of the debunking arguments that I described, the

---

[6] For instance, she writes, "Barring such a coincidence, the only conclusion remaining is that many or most of our evaluative judgments are off track" (121–122), and "we may understand these evolutionary causes as having tracked the truth; we may understand the relation in question to be a tracking relation" (Street 2006, 125).

[7] For instance, Enoch writes, "On the other horn of the dilemma, the only realist-friendly explanation of the correlation seems to be a tracking account of some sort, according to which our normative judgments and the evolutionary pressures shaping them causally track the independent normative truths" (Enoch 2010, 426; my emphasis); Copp talks about beliefs as truth tracking: "Realists must either affirm or deny a thesis I will call *the tracking thesis*, the thesis that natural selection so affected our psychology that our moral beliefs tend to track the moral facts" (Copp 2008, 191). Griffiths and Wilkins also talk about evolution as a truth-tracking process.

epistemic problem can be understood in terms of truth tracking—as failure of adherence or sensitivity. Thinking about the problems posed by etiological debunking arguments as a matter of how information about causes influences our assessment of whether we track the truth of certain propositions helps clarify the relation between the different types of problems that etiological debunking arguments seem to pose.

I will begin by presenting Sherrilyn Roush's account of truth tracking, which I adopt, though I do not use her account of knowledge as truth tracking in this project. Then, I discuss the value of possessing true beliefs and lacking false beliefs about certain propositions over time in practical contexts where we have goals that we want to achieve. If it is valuable for accomplishing our goals to have true beliefs and lack false beliefs about these propositions over some duration, I argue, it is also valuable to track the truth of these propositions over that time period. I extend these points into the moral context, adopting the assumption that we have the goal of accomplishing any moral obligations we may have. I consider several objections, and then I compare the value of truth tracking in practical contexts with some of the other epistemic properties that we might like our beliefs to have. Finally, I discuss the relationship between truth tracking and the causes of our beliefs and argue that information about the causes of beliefs can shed light on whether those beliefs track the truth.

## 2.1 ROUSH'S ACCOUNT OF TRUTH TRACKING

In *Tracking Truth*, Roush provides a truth tracking account of knowledge (2005). She modifies Robert Nozick's (1981) counterfactual truth tracking account of knowledge by presenting the conditions for truth tracking as conditional probabilities and adding a clause that preserves

closure.[8] On her account, an agent tracks the truth of a proposition if and only if (1) the probability that the agent does not believe *p* conditional on *p* being false meets some high threshold and (2) the probability that the agent believes *p* conditional on *p* being true meets some high threshold. Condition (1) is known as the sensitivity condition. Condition (2) is known as the adherence condition. An agent can meet the adherence condition for some proposition but fail to meet the sensitivity condition for that proposition, and vice versa. On Roush's account it is sufficient for the agent to know *p* if the agent believes *p*, *p* is true, and the agent tracks the truth of *p*. To know in this way is to know by tracking the truth of *p*.[9]

---

[8] The simplest version of Nozick's original account of knowledge said that a subject *S* knows that *p* if and only if (1) *p* is true, (2) *S* believes that *p*, (3) if *p* were not true, *S* would not believe that *p*, and (4) if *p* were true, *S* would believe that *p* (Nozick 1981, 172-178). Conditions (3) and (4) are the truth tracking conditions. They are subjunctive conditionals that can be understood in possible worlds semantics as follows: (3) in the closest possible world where *p* is not true, *S* does not believe that *p*, and (4) in the closest possible world (or in a few of the closest possible worlds) where *p* is true, *S* believes that *p*. Nozick dealt with various counterexamples by modifying (3) and (4) so that they were fixed to a particular method (e.g. "(3) If p were not true and S were to use [method] M to arrive at a belief whether (or not) p, then S wouldn't believe, via M, that p" (179). One consequence of Nozick's account was that it denied that knowledge is closed under known entailment.

[9] On Roush's account, the agent may alternately know that *p* by known implication, if and only if the following conditions hold: *p* is true, the agent believes that *p*, and the agent knows via

Although it is not part of her account of knowledge, Roush also discusses the property of *safety*, which others have suggested may be a necessary condition for knowledge (Sosa 1996, 2000). Roush contrasts safety with sensitivity in particular, arguing that sensitivity is a more valuable epistemic property for a variety of reasons. Like sensitivity and adherence, safety was originally introduced as a counterfactual condition, but Roush presents it as the conditional probability that *p* is true given that the agent believes *p*; for a belief to be safe this probability should meet some high threshold.[10] There is a fourth epistemic property that one might consider along with sensitivity, adherence, and safety. This property, which I will refer to as *safety counterpart*, is a high probability that *p* is false given that the agent does not believe *p*. This epistemic property is not of much interest for questions related to knowledge, but it may yet be a desirable epistemic property, and sensitivity, safety, adherence, and safety counterpart are interdefined, so its existence is worth noting.

Roush trades counterfactuals for conditional probabilities in part because Nozick's counterfactuals required only that the subject would believe *p* in one or a few closest possible worlds where *p* is true (in the adherence condition) or would not believe p in one or a few closest possible worlds where *p* is false (in the sensitivity condition) (28). Considering what the subject would believe in only one or a few nearby worlds makes his account of knowledge too weak, producing a cluster of counterexamples. More importantly for my purposes, if we are interested

---

tracking some set of propositions that imply *p*, and the agent knows that this set of propositions implies *p*.

[10] Sosa presented safety as the following condition: If *S* were to believe *p*, *p* would be true (1999, 146). Or, in all nearby worlds where *S* believes *p* (via method *M*), *p* is true.

in whether a subject is likely to believe a proposition under the range of circumstances that the subject may face in the course of goal completion, it is not adequate to consider only how the subject's attitude would vary given one or a few variations in the circumstances. Roush observes that it may be possible to develop a theory of subjunctive counterfactuals that requires for their evaluation that one consider more worlds than just the closest or a few of the closest worlds. However, no such theory is currently available. By contrast, assessing the adherence and sensitivity conditions in the form of Roush's conditional probabilities naturally involves examining a variety of ways the world might have been different (28-29).

## 2.2 THE PRACTICAL VALUE OF TRUE BELIEF, ABSENCE OF FALSE BELIEF, AND TRUTH TRACKING

Roush offers two lines of argument in support of the claim that truth tracking is valuable. First, when true belief (and absence of false belief) is *useful*, then truth tracking is valuable, because meeting each of the truth tracking conditions for some proposition means that the probability is high that we will have a true belief and lack a false belief in the future and in different contexts. This helps us *do* things; it gives us power to control the world (26). Second, Roush argues that truth tracking may be intrinsically valuable. In the moral domain, where we are concerned with completing any obligations we might have, what should interest us is the practical value of true belief and truth tracking—the value of each of these for accomplishing obligations. So, I will set aside questions about the intrinsic value of both true beliefs and truth tracking.

In this section, I will discuss the practical value of true belief and absence of false belief about certain propositions that is conferred by having some goal. For an agent with a goal, there

is a set of propositions with the feature that believing a proposition in this set when that proposition is true and lacking a belief in it when it is false increases the agent's chances of accomplishing her goal. In the next section I will argue that when possession of true belief and absence of false belief is valuable for practical reasons, this makes truth *tracking* a valuable epistemic property to have. Then I will apply these points to contexts in which one has specifically moral goals.

I will introduce some shorthand to talk more easily about the idea that it is useful to have true beliefs and to lack false beliefs about certain propositions. An agent's attitude about a proposition *matches* the truth value of the proposition if it is the case that (1) *p* is true and the agent believes *p*, or if it is the case that (2) *p* is false and the agent does not believe *p*.[11] We can call the first situation *positive matching*, and the second, *negative matching*. So, for the proposition "There's a hummingbird at the window," the agent's attitude matches this proposition if either of these two holds: (1) there is in fact a hummingbird at the window and the agent believes it, or (2) there is no hummingbird at the window, and the agent does not believe that there is, whether because the agent has no belief on the subject or believes that there is no hummingbird at the window. So, my first claim is that for any agent with a goal, there is a set of propositions with the feature that the agent's possession of attitudes that match the truth value of

[11] Matching in this sense should be distinguished from its use in a correspondence theory of truth on which a proposition is true if it matches the world. My thesis is independent of one's theory of truth. I should note that Roush appears to use the term "matching" in this way a few times in her (2006).

those propositions increases the agent's chances of completing the goal.[12]

For instance, suppose that the agent's goal is to drink a cup of coffee within the next fifteen minutes. There are a number of propositions such that if the agent's attitude about those propositions matches their truth value, against some set of ordinary background beliefs, this increases the agent's chances of completing his goal. These might include: (a) that the agent is at home, with no coffee shops nearby; (b) that there is a coffee maker on the kitchen counter; (c) that coffee is stored on the top shelf; (d) that all the coffee mugs are dirty, etc. If it is false that the coffee is stored on the top shelf, it would be better for the agent to lack a belief in this proposition than to believe it. Likewise, it would be better for the agent to believe that the coffee maker is on the counter than not believe it, if the coffee maker is indeed on the counter.

For some of these propositions it is valuable to have an attitude about them that matches their truth value during specific times over the course of goal completion. If it is true that all the coffee mugs are dirty, it is useful for the agent to believe this while he still has time to wash a mug—say, up to fourteen minutes in, to leave him one minute to wash the mug. Up to fourteen minutes in, it is useful for him to have an attitude about the proposition expressed by the sentence that "all the coffee mugs are dirty" that matches the truth value of that proposition. It would be best, with regard to any of these propositions, for the agent to have attitudes that matched their truth values at all the times that mattered for goal completion.

To give another example—a classic case in which philosophers have discussed the practical value of true belief is the conversation between Socrates and Meno about the road to

---

[12] It is nonstandard to refer to withholding belief as an attitude, but there is no natural alternative to refer to these three relationships between S and *p*.

Larissa. We can understand *getting to Larissa* to be the goal or the relevant action in this case. We can imagine some relevant propositions, such as "The road I am on is the road to Larissa" and "To get to Larissa, if I am at a fork, I should take the lefthand path." One's attitudes about these propositions at various times would affect one's chances of reaching Larissa. It would be ideal to have attitudes that match the truth value of the first proposition over the course of the entire trip. For the second proposition, it would probably be sufficient to have attitudes matching the truth value of this proposition only on those occasions when one finds oneself at a fork.

These are cases of goals—coffee making and traveling to a destination—from everyday life. But the point extends to other goals, such as goals in science, engineering, politics, economics, etc., and, as I will discuss in section 4, to goals in the moral domain. So, for each goal there is a set of propositions such that having attitudes about those propositions that match their truth value (at the right times) increases the agent's chances of goal completion. I will refer to this set of propositions as "relevant propositions."

Of course, there are complications. For some goals there may be disjunctions of propositions or of sets of propositions, such that possessing attitudes that matched the truth values of all the propositions in any one of those sets would make the agent more likely to complete the goal. In fact, this complexity will probably apply in most cases. For instance, if there were three equally good recipes for making oatmeal cookies, a person would be equally well off if he had matching attitudes about what to do at each step for any one of these recipes. In such a case, each set of propositions would be relevant for determining the agent's probability of goal completion. Furthermore, in some cases there will be a set of propositions, such that if the agent has attitudes that match the truth value of all the propositions in that set, the agent's chances of goal completion increase, but if the agent has attitudes that match the truth value of

only some of the propositions, then there is no (or little) influence on the chances of goal completion. For instance, in the coffee case, if the agent has attitudes that match propositions (a)-(d) above but the agent has no attitudes that match the truth values of propositions related to operating the coffee maker, the fact that the agent matches propositions (a)-(d) does little to increase the agent's chances of goal completion. Nonetheless, I think for a given goal we will usually be able to generate a list of propositions or sets of propositions with the feature that matching makes a significant difference to one's chances of goal completion among many or the most common sets of background beliefs. I will address some of the challenges associated with this task in section 6.

When one sets out to accomplish a goal, there is some probability that one's attitudes about each of the relevant propositions will match the truth values of the propositions at the relevant times. Since having attitudes that match the truth values of these propositions increase one's chances of goal completion, a higher probability of matching means a higher probability of goal completion. As a result, when the agent sets out to accomplish the goal, it would be best if the probability is high that the agent's attitudes regarding the relevant propositions match the truth values of those propositions at all the relevant times: the agent would surely want this probability to be high, and, if it were an option, the agent would want to increase this probability. So, in cases where it is pragmatically valuable to possess attitudes that match the truth value of a proposition, it is also (instrumentally) valuable for the agent to have a high probability of matching when she sets out to accomplish her goal.

## 2.3 THE RELATIONSHIP BETWEEN MATCHING, TRUTH TRACKING, AND SAFETY, AND HOW TO INVESTIGATE THE PROBABILITY OF MATCHING

As described above, the agent's attitude about a proposition, *p*, matches the truth value of *p* if either (1) the agent believes *p* and *p* is true (positive matching) or (2) the agent does not believe *p* and *p* is false (negative matching). To put it formally, let *Bp* stand for the event that the agent believes that *p*, and *M* for the event that the agent's attitude matches the truth value of *p*. Then, we can write the definition of matching as follows:

$$Pr(M) = Pr[(Bp\&p)v(\sim Bp\&\sim p)].$$

Bp&p and ~Bp&~p are mutually exclusive, so this definition implies:

$$Pr(M) = Pr(Bp\&p) + Pr(\sim Bp\&\sim p) \qquad (1)$$

$$Pr(M) = Pr(p\&Bp) + Pr(\sim p\&\sim Bp) \qquad (2)$$

Applying the definition of conditional probability to each term on the right hand side of Eq. (1), we get

$$Pr(M) = Pr(Bp|p)Pr(p) + Pr(\sim Bp|\sim p)Pr(\sim p). \quad (3)$$

Note that Pr(Bp|p) is adherence and Pr(~Bp|~p) is sensitivity.

Now, if we apply the definition of conditional probability to Eq. (2), we get

$$Pr(M) = Pr(p|Bp)Pr(Bp) + Pr(\sim p|\sim Bp)Pr(\sim Bp) \qquad (4)$$

where Pr(p|Bp) is safety and Pr(~p|~Bp) is safety counterpart.

Consequently, one may investigate the probability of matching in a variety of ways, depending on the information available. If we already know Pr(Bp&p) and Pr(~Bp&~p), we need not use Eq. (3) or (4). If we do not know these probabilities, we can use Eq. (3) or (4). If we are more likely to have information about Pr(p), adherence, and sensitivity we can use Eq. (3); if we are more likely to have information about Pr(Bp), safety, and safety counterpart, we can use

Eq. (4).

Eq. (3) suggests that Pr(M) depends on both adherence and sensitivity just in case 0 < Pr(p) < 1. Otherwise, one of the terms on the right hand side of Eq. (3) is zero, and Pr(M) depends only on adherence or sensitivity. Eq. (4) suggests that Pr(M) depends on both safety and safety counterpart just in case 0 < Pr(Bp) < 1. Otherwise, one of the terms on the right hand side of Eq. (4) is zero, and Pr(M) depends only on safety or safety counterpart.

Furthermore, we can apply Bayes' Rule to adherence, sensitivity, safety, and the safety counterpart:

$$[\text{Adherence}]\ \Pr(Bp \mid p) = \frac{\Pr(p \mid Bp)\Pr(Bp)}{\Pr(p \mid Bp)\Pr(Bp) + \Pr(p \mid {\sim}Bp)\Pr({\sim}Bp)}$$

$$[\text{Sensitivity}]\ \Pr({\sim}Bp \mid {\sim}p) = \frac{\Pr({\sim}p \mid {\sim}Bp)\Pr({\sim}Bp)}{\Pr({\sim}p \mid {\sim}Bp)\Pr({\sim}Bp) + \Pr({\sim}p \mid Bp)\Pr(Bp)}$$

$$[\text{Safety}]\ \Pr(p \mid Bp) = \frac{\Pr(Bp \mid p)\Pr(p)}{\Pr(Bp \mid p)\Pr(p) + \Pr(Bp \mid {\sim}p)\Pr({\sim}p)}$$

$$[\text{Safety counterpart}]\ \Pr({\sim}p \mid {\sim}Bp) = \frac{\Pr({\sim}Bp \mid {\sim}p)\Pr({\sim}p)}{\Pr({\sim}Bp \mid {\sim}p)\Pr({\sim}p) + \Pr({\sim}Bp \mid p)\Pr(p)}$$

Note that adherence and sensitivity are defined in terms of safety and safety counterpart, and, in turn, safety and safety counterpart are defined in terms of adherence and sensitivity.[13] Importantly, this means that adherence, sensitivity, safety, and safety counterpart are closely

[13] For instance, adherence = safety x Pr(Bp) / (safety x Pr(Bp) + Pr(1-safety counterpart) x Pr(~Bp)).

connected. If one fixes Pr(Bp), Pr(p), and two of these four conditional probabilities, one can determine the two remaining conditional probabilities.

Eq. (3) suggests that if Pr(p) = 1, then Pr(M) is determined by adherence. But if Pr(p) = 1, then Pr(p|Bp) = 1 and Pr(~p|~Bp) = 0. Adherence is thus determined by Pr(Bp) alone, and so is Pr(M). Hence, if Pr(p) = 1, we should be interested only in Pr(Bp). (Also, note that if Pr(p) = 1, Pr(~Bp|~p) is undefined). Likewise, Eq. (3) suggests that if Pr(~p) = 1, then Pr(M) is determined by sensitivity. But if Pr(~p) = 1, then Pr(p|Bp) = 0 and Pr(~p|~Bp) = 1. Sensitivity is thus determined by Pr(~Bp) alone, and so is Pr(M). Hence, if Pr(~p) = 1, we should be interested only in Pr(~Bp).

Similarly, suppose there were some propositions that we were bound to believe: Pr(Bp) = 1. Then Eq. (4) suggests that Pr(M) is determined by safety. But if Pr(Bp) = 1, then Pr(Bp|p) = 1 and Pr(~Bp|~p) = 0. Safety is thus determined by Pr(p) alone, and so is Pr(M). Hence, if Pr(Bp) = 1, we should be interested only in Pr(p). Also, suppose there were some propositions that we were bound not to believe: Pr(~Bp) = 1. Then Eq. (4) suggests that Pr(M) is determined by the safety counterpart. But if Pr(~Bp) = 1, then Pr(Bp|p) = 0 and Pr(~Bp|~p) = 1. The safety counterpart is thus determined by Pr(~p) alone, and so is Pr(M). Hence, if Pr(~Bp) = 1, we should be interested only in Pr(~p).

## 2.4 THE PRACTICAL VALUE OF MATCHING AND TRUTH TRACKING IN THE MORAL CONTEXT

The argument about the pragmatic value of matching and tracking certain propositions for the purposes of completing goals can be extended to the moral context. For the purposes of

accomplishing moral goals, there are propositions with the feature that matching those propositions increases the agent's chances of completing the moral goal. For instance, very broadly, a person might have the goal of accomplishing his moral obligations—in some particular context, or over the course of his life. Or a person might have the goal of becoming a virtuous person, or exhibiting a particular virtue, such as bravery. For the purposes of accomplishing these goals, there are sets of propositions—including certain moral propositions—with the feature that if the person has a matching attitude about these propositions, this increases the chances that the person will accomplish his goal. Among the set of propositions relevant for any given moral goal will be certain moral propositions: the proposition that one has the some moral obligation, propositions about the nature of the moral obligation and means for achieving it, propositions about which features of a situation are morally relevant, and (possibly) metaethical propositions. As the person sets out to accomplish the goal, the person will want to have a high probability of matching these propositions—it will be better for the person to be in a position where he has a high probability of matching these propositions.[14]

For instance, suppose an agent has the goal of acting properly toward strangers. The agent is sitting on a bus and another person is standing in front of her. Suppose that the fact of

---

[14] It may be that it would be just as good not to have matching attitudes about the truth of the proposition but instead matching pro- and anti-attitudes at the appropriate times or simply behaviors that match one's obligations. Provided that there are appropriate and inappropriate pro- and anti-attitudes and behaviors, it would then still be possible for one to track or fail to track something even if it is not truth. This might mean that my account could be adapted to apply to some non-cognitivist views. I plan to explore this possibility in the future.

the matter is that to act properly toward strangers, the agent should give up her seat to the stranger in front of her. The agent will be more likely to accomplish her goal in this situation if she has a high probability of matching certain propositions over the course of the bus ride. It is better for the agent if she has a habit of staying attuned to her surroundings in a way that makes it likely that she will acquire and maintain matching attitudes about the (nonmoral) proposition that there is someone standing in front of her. The agent will also be more likely to achieve her goal if she can form matching attitudes about (nonmoral) propositions related to the frailty level of the person in front of her. Matching attitudes about that sort of proposition, though, will only increase her chances of goal completion if she also has a matching attitude about the relevance of frailty to the question of how to act properly to strangers. Consider the moral proposition that when one is sitting on a crowded bus and a more frail person is standing, the way to act properly to strangers is to give up one's seat. It will increase the agent's chances of completing her goal of acting properly toward strangers if she is likely to believe this moral proposition, given that the proposition is true. Similarly, consider the moral proposition that the proper treatment of strangers is independent of the commands of the bus driver. The agent will be better off if she is likely to match the truth value of this proposition. If the agent is inclined to believe that the proper treatment of strangers is independent of the commands of the bus driver when this proposition is true, she will be more likely to complete her obligation. For example, if the bus driver falsely instructed her that frailty is morally irrelevant on his bus, she will be more likely to reject his claims.

Before moving on, I want to discuss what it means for my account if some moral propositions relevant for goal completion are necessarily true. Accounts of knowledge that employ sensitivity and adherence or safety have had trouble dealing with necessary propositions.

There are many necessary propositions that people are thought to know. Yet sensitivity is undefined in the cases of propositions that are necessarily true, as noted above, and so if one makes sensitivity a necessary condition for knowledge, true beliefs in necessary propositions will not count as knowledge. The truth tracking theorist could make an exception for necessary propositions, and say that beliefs in necessary propositions only need to meet the adherence condition to count as knowledge. However, this will result in some cases of belief in necessary propositions that epistemologists think are not knowledge getting counted as knowledge. For instance, suppose one has a belief in a necessary proposition. If one has a high probability of believing the necessary proposition solely because a benevolent demon ensures one's belief, one will have a high degree of adherence. But many epistemologists will not count this as a case of knowledge. Similarly, many epistemologists think that safety is fulfilled too easily in cases of necessary propositions—essentially, on an account where a true belief's safety suffices for knowledge, one knows a necessary proposition whenever one believes it, which will lead us to categorize too many cases of true belief as knowledge.

However, because what motivates this project is concern for accomplishing any obligations we might have, and because, as I have argued, having certain matching attitudes will likely make a substantial difference in whether we do this, I am not pursuing an account of knowledge here, nor assessing whether we have moral knowledge. For my task, the possibility that some moral propositions are necessary propositions pose no special problem. I am investigating the probability that we will match various moral propositions relevant for our moral goals. I noted earlier that if $Pr(p) = 1$, as in the case of necessary truths, our probability of matching is determined solely by adherence, which in turn is determined solely by $Pr(Bp)$. If for an agent with a goal, possession of attitudes that match the truth value of certain necessary

propositions increases the agent's chances of goal completion, this just means that the agent will be better off if she has a high Pr(Bp) during the time relevant for goal completion. That a moral proposition is necessary does not undermine the possibility that it is valuable for the agent to have attitudes that match the truth value of the proposition, and to have a high probability of matching.

Admittedly, if all the moral propositions that are relevant for goal completion are necessary, this will mean that none of the other probabilities discussed earlier—Pr(p), adherence, sensitivity, safety, and safety counterpart—matter for our chances of matching relevant moral propositions. But there are some moral propositions relevant for goal completion that are not necessary truths, and so these other probabilities do play a role in determining our chances of matching relevant propositions, and, as a result, they bear investigating.

The relevant contingent moral propositions are moral propositions such as those expressed by the following sentences: (1) Barack Obama should give some money to a charity today; or (2) Barack Obama should give some money to Oxfam this year. Suppose (1) expresses this proposition: Barack Obama should give some money to a charity on May 16, 2013. For evaluating the truth of this proposition, the only worlds that are relevant are those in which Barack Obama exist, it is May 16, 2013, and Obama is capable of giving to a charity. Now, within this set of worlds, many things vary, including facts about Obama and facts about charities. In some worlds Obama would suffer greatly if he gave to a charity; in some worlds, all charities are ineffective. Now, suppose (2) expresses the proposition that Obama should give some money to Oxfam in 2013. The relevant worlds would be those where Obama exist, Oxfam exists, it is 2013, and Obama is capable of giving some money to Oxfam. Within these worlds, facts about Obama and facts about Oxfam would vary across worlds. In some worlds, Oxfam

might do more bad than good.

Suppose our goal is to complete our moral obligations.[15] Many of the propositions that will increase our chances of completing this goal have to do with what our obligations are or what the obligation of some set of people is at some time or under some circumstances. The person or set of people that appear in the proposition and the time or the set of circumstances that appear in the proposition restrict which worlds are relevant. But within the set of relevant worlds, the proposition will be true in some and false in others, depending on facts that vary between worlds. Suppose Jose finds himself seated on a bus at time *T*, and thinks to himself that he should give up his seat to someone else. This proposition is contingent: In some worlds at *T*, such as worlds where everyone on the bus is more frail than Jose, Jose's obligation will be give up his seat on a bus; in other worlds at *T*, such as worlds where the only others on the bus are husky football players, Jose's obligation would be to retain his seat on the bus. (This example assumes that frailty is morally relevant). Consequently, it would be valuable for Jose to have a high degree of sensitivity as well as adherence with regard to this proposition. For instance, it would be good if Jose would have acquired this belief if everyone around him were frail, and if he would have not held this belief if he were surrounded by football players.

Is there something tricky going on here when I appeal to propositions like the proposition that I should not give up my seat to this person? Roush has used a similar sort of proposition as an example of a (nonmoral) proposition to which truth tracking conditions are applied. She talks about the proposition that there is a tiger in front of a subject (270), noting that this could be true

---

[15] If we have no moral obligations, then this goal is not possible, which means my account should not be applied to this case.

or false: "The truth value of *p* may change; the tiger may be gone today but back tomorrow" (Roush "Value of knowledge," 259). In some worlds, there is a tiger in the vicinity of the agent in question; in other worlds, there is no tiger in the vicinity of the agent in question. Similar change in truth value may occur in the bus case. In some possible worlds in which the agent in question is sitting on the bus, it is true that the agent should give up her seat to the person in front of her; in other worlds it is false. Supposing, then, that there are contingent moral propositions with the feature that matching those propositions increases our chances of goal completion, then Pr(Bp) will not be the only probability that matters for our chances of matching; sensitivity, adherence, safety, safety counterpart and Pr(p) may also be relevant.

## 2.5 PRELIMINARY WORRIES

To clarify how my account is meant to work, I will address some preliminary worries about the claim that for any given goal there is a set of propositions such that if the agent has attitudes toward these propositions that match their truth value, this makes some significant difference to goal completion.

First, one might wonder whether for any given goal there might be a vast number of potentially relevant propositions. As a result, one might ask how feasible it is to identify the relevant propositions with the feature that matching them would have a significant positive effect on one's chances of goal completion (against some set of background beliefs), and how one would go about doing this. I acknowledge the possibility that a vast number of propositions may be relevant for a goal, but I think that for each goal we can find some propositions that are clearly more relevant than others and some propositions that are clearly less relevant. For

example, in the coffee case, propositions about very improbable events will be correspondingly less relevant for the chances of goal completion. Propositions about an earthquake knocking all the coffee onto the floor, for instance, will not be terribly important to track. In other cases, when one's goal has to do with certain improbable events, propositions about those improbable events will be clearly relevant. For instance, if one's goal is to construct a skyscraper that will withstand earthquakes, it will important to track propositions about the frequency of earthquakes and building methods effective for preventing damage resulting from these improbable events. So, one thing to say is that some of the propositions will be more important to match than others. There is no general rule for how one would go about the task of identifying relevant propositions for all goals. The most relevant propositions for bridge building differ from the most relevant propositions for getting through the day. However, in certain domains there may be propositions that are usually relevant for goals in that domain. In engineering, it is surely valuable to have matching attitudes about sets of mathematical propositions and principles from physics. Similarly, in the domain of morality it may be generally useful to match certain propositions, such as propositions about moral realism, authority independence, truth makers of moral rules, etc. Other propositions will likely be relevant in most cases, such logical and basic mathematical propositions. At the same time, many of these propositions that are generally valuable to track will not bear investigating—e.g. because we are already confident that the person in question tracks those facts.

To expand on the moral case, suppose that we have the moral goal to complete any obligations that we might have. What propositions are relevant for this goal, in the sense that having an attitude about this proposition that matches increases ones chances of goal completion? It is harder to determine which propositions are relevant if we do not make

assumptions about the nature of our moral obligations, but I suggest that generally it is good to match propositions that have to do with *the content of one's moral obligations and the means for achieving them*. It may also be good to match metaethical propositions about the nature of obligations (such as what makes moral propositions true) and propositions about the reasons an action ought to be done in a particular case. For instance, suppose that an egoist believes (*p*) he ought to refrain from killing his friend in some circumstance—suppose this is true, that this is his obligation in that circumstance—and he believes (*q*) he should do this because he would face punishment otherwise—which is false; this is not the reason he should abstain from killing. His non-egoist peer possesses the first belief but not the second; the non-egoist believes (*p*) he should not kill his friend in the relevant circumstances and (*r*) *p* because murder is always wrong. Each has a true belief about his obligation: he should not kill his friend in these circumstances. But the egoist would be better off if he lacked his false belief in *q* and possessed a true belief in *r*. The egoist does not track the truth of this proposition as well as the non-egoist, because if the situation were slightly different such that the egoist would not be punished if he killed his friend, the egoist would no longer believe *p*. By contrast, the non-egoist would retain his belief that *p*.

Second, one might worry that one's attitudes about propositions might not make a significant difference to one's chances of goal completion in comparison with other things in the world, such as physical facts, the truth value of the propositions themselves, etc. For instance, one's chances of making coffee in the next fifteen minutes are more influenced by whether there is coffee close enough that one could access it in fifteen minutes than by whether one has a matching attitude about whether there is coffee close by—if there is no coffee nearby, it hardly matters what attitude one has about whether any coffee is accessible. This raises the worry that perhaps to assess our chances of goal completion, we should look at factors other than our

attitudes about propositions and whether they match.

I have three things to say about this. First, in *some* cases, whether our attitudes match certain propositions clearly matters a great deal. For instance, when our goal is to build a stable bridge, our chances are heavily influenced by our attitudes about propositions. Second, even if matching propositions has just a small influence over our chances of goal completion, we will still want to investigate our chances of matching in cases where our attitudes are the only things that we have control over. Third, in some of the cases in which any contributions matching attitudes make are swamped by other factors, this is because outside factors make goal completion impossible to begin with. I mean to exclude these cases, where there is actually no chance of goal completion, from my story. When I talk about a goal, I am talking about a task that it is possible for the agent to complete.

## 2.6 OBJECTIONS

Some have challenged the claim that possession of true belief is generally pragmatically valuable. This point is not exactly a challenge to my thesis, which is not that matching is generally pragmatically valuable but that for any given goal there are some propositions such that it is pragmatically valuable to have attitudes that match the truth values of those propositions. The cases offered by proponents of this challenge raise the worry, though, that whether it is valuable to possess attitudes that match the truth value of a particular proposition may depend significantly on context, and so determining whether it would be valuable to possess attitudes that match the truth value of some proposition may be somewhat more difficult than one might have thought.

There are certainly some propositions, in some contexts, where it would be better for goal achievement to have an attitude that does not match the truth value of the proposition—it would be better to lack belief in the proposition when it is true or to believe the proposition when it is false. Stich offers an example of a situation in which it would be better to lack a true belief about a certain proposition: Harry believes his flight leaves at 7:45 a.m.; he gets on the plane, it crashes, and he dies. For this case, we might assume that Harry's goal is survival over the course of the day. Adding a true belief to the agent's body of beliefs worsens his chances of survival. By contrast, had he falsely believed that the plane left at 8:45, he would have avoided the crash and lived (Stich 1990, 123). Stich concludes that more true beliefs are not always better than fewer, and, we can add, sometimes it is better to possess a false belief in a proposition than not. This is surely true.

Background beliefs and incomplete sets of beliefs seem to contribute a great deal to whether one has a problem in cases like the one Stich raises. Here Harry has an incomplete set of beliefs about the situation, and some of his other beliefs are false, as Stich notes. If Harry had the true belief that the flight leaves at 7:45, but also the true belief that the plane would crash, he would not have a problem. The fact that it is better not to match the proposition about the time the plane leaves only comes about because Harry fails to match this other proposition, that the plane is going to crash. For my purposes, this case just underlines the fact that sometimes whether it is valuable to match a proposition is affected by one's attitudes about certain other propositions. If Harry matches the proposition that the plane is going to crash, then it is good to match certain propositions; if he does not match this key proposition, then it will be better to fail to match certain of these propositions. These sorts of complications will make it trickier to identify the propositions with the feature that matching makes a significant positive difference to

the chances of goal completion. It might be necessary to look instead for the propositions with the feature that matching makes a significant positive difference to the chances of goal completion *against a certain background of other beliefs*.

Furthermore, possession of the false belief that the flight leaves at 8:45 would allow Harry to achieve his goal in the particular case described, but goal achievement here would happen in a sense by accident, because the case described is also very improbable. Normally, it would be advantageous for Harry to having a matching attitude about the proposition that his flight leaves at a particular time. After the fact, we can say in this case that it would have been better for Harry's chances of goal completion if he had not match the proposition that the flight leaves at 8:45. But beforehand, because it is improbable that the plane will crash, what is best for Harry's chances of goal completion is to match this proposition: to not believe that the flight leaves at 8:45, and to have the true belief that the flight leaves at 7:45.

One might also ask whether, for a proposition whose matching seems likely affect our chances of goal completion, it is so critical that we believe that specific proposition when it is true and lack belief in it when it is false, or whether we could do just as well, or nearly as well, by believing a similar (but strictly false) proposition *q* when *p* is true and lacking belief in *p* and propositions similar to *p*, such as *q*, when *p* is false. Practically, this might mean that when we find a proposition that we suspect would make a difference to goal completion, and we find that we have a low probability of matching this proposition, we might also want to check whether there are similar propositions that we tend to believe and disbelieve at the relevant times. If so, in these cases, failure to match *p* will not be so terrible for our chances of goal completion. For instance, in the coffee case, it might be useful for the agent to have an attitude that matches the true proposition *(p*) that there is coffee stored in the cupboard. Perhaps the agent does not match

this proposition, because he does not realize that the wooden box on the wall is a cupboard and instead thinks it is a bookshelf, but whenever *p* is true he tends to believe the false proposition *(q*) that there is coffee stored in the bookshelf on the wall. Believing *q* when *p* is true and not believing *q* when *p* is false would presumably be good enough for the purposes of accomplishing his goal of drinking coffee in the next fifteen minutes. So, if I conclude that our chances of completing some moral goal are compromised by the fact that we do not appear to track some proposition seemed to be relevant for goal completion, one might object by identifying propositions similar to *p* that we appear to "track" in place of *p*, and contending that we are not as badly off as I have claimed. This is an important worry, and it will have to be taken into consideration when identifying the propositions that are relevant for one's moral goal.

## 2.7 OTHER EPISTEMIC PROPERTIES

One might ask whether various other epistemic properties are more pragmatically valuable than (or at least as pragmatically valuable as) the property of having a high probability of matching a proposition. Perhaps there are other epistemic desiderata that it would be better to investigate for the purposes of assessing our epistemic situation in relation to goal completion. I will modify Alston's helpful categorization of epistemic desiderata as a way to organize the desiderata that might comprise an alternative to high probability of matching (Alston 2005). First, there is the property of a belief being true. Second, there are properties that are "truth-conducive." Sensitivity, adherence, and safety fit into this category. We can also count the following properties of belief as falling into this category: being supported by adequate evidence, being the product of a reliable process, being the product of properly functioning cognitive processes, and

being the product of exertion of an intellectual virtue. These properties fall into this category to the extent that they render the belief likely to be true. Third, there are properties that Alston designates as deontological features of belief. We can put the properties of being justified and known into a fourth category—these are properties whose nature is highly disputed. A fifth category consists of properties of systems of beliefs, such as coherence.

Respect for the value of truth does not compete with my argument for the value of matching and the probability of matching; it is part of my story, since one matches a proposition either when one has a true belief about the proposition or when one lacks a false belief about it. It is the value of matching a proposition that makes a high probability of matching valuable. I have argued that inasmuch as we are interested in true beliefs or matching, we will be interested in the probability of matching, and thus in the properties of sensitivity, adherence, and safety. To assess our epistemic position for the purposes of assessing and improving chances of goal completion, though, we cannot just look at truth or matching alone. Because it would be useful for us to match a proposition in the future, we are better off in current position if we have a high probability of matching at those future moments than if we do not. So, although in the end we will either match the proposition or not at each of the relevant moments, we do not go about investigating our epistemic position just by assessing whether we match at certain moments (the only moments we'd be able to do this for would be current and past moments); we also have to assess whether we are *likely* to match at relevant moments in the future.

Truth conducive properties are valuable because they increase the probability that one's attitude about a proposition matches the truth value of that proposition. For instance, if a belief is the product of a reliable process, this means that the belief is likely to be true. In other words, given that you believe *p* via this reliable process, the probability of *p* is high (assuming that the

process in question is the only way that you would have acquired the belief). The properties of being supported by adequate evidence, being the product of properly functioning cognitive processes, and being the product of exertion of an intellectual virtue, too, are pragmatically valuable in my view because they increase the probability of matching. So, we can use information about the evidence for a belief, the reliability or proper functioning of the process from which it came, or whether its production involved the exertion of a virtue, to gauge how likely one is to match that proposition. But these properties do not confer pragmatic value apart from inasmuch as they increase the probability of matching.[16]

Deontological features of belief are usually claimed to contribute non-instrumental value. For instance, that a belief is held permissibly, that it is formed and held responsibly, that its origins do "not contain violations of intellectual obligations" (Alston 2005, 45) are meant to be intrinsically valuable or at least not valuable for the purposes of goal completion. So, if we wanted to investigate our epistemic status not with regard to goal completion but something else, or if we wanted to investigate the non-pragmatic value of our beliefs, we might look at these properties. But since my task has to do with the pragmatic value of beliefs, these properties, if their value is not pragmatic, are not competitors with the property of having a high chance of matching.

I have been arguing for the value of truth tracking and high probability of matching without weighing in on its relation to knowledge, but of course, the conditions of sensitivity, adherence, and safety were introduced as conditions for knowledge. A natural question is the following: to assess our chances of completing our moral goals, why not investigate whether we

[16] Roush argues something like this in her (2010).

have moral knowledge? I am not investigating whether we have moral knowledge because whether knowledge does anything to improve our chances of goal completion depends on which properties compose or are necessary for it. I do not need to weigh in on the disputed question of the nature of knowledge in order to show that truth tracking is the property that should concern us when we investigate the epistemic status of our moral beliefs. However, if the conditions of sensitivity, adherence, or safety are related to knowledge in some way, the pragmatic value conferred by these conditions can help explain the pragmatic value of knowledge.[17] Knowledge may be in some nonpragmatic sense have greater value than truth tracking belief, but even if this is true, having knowledge instead of truth tracking belief would not help us accomplish obligations, which is my primary focus in this project.

The criteria for justification, like the criteria for knowledge, are disputed. I am not investigating our chances of completing moral goals by looking at whether our moral beliefs are justified for the same reasons I am not looking at whether our moral beliefs are known: whether justification is pragmatically valuable will depend on what it is, and it seems more efficient to discuss the value of such properties directly. If justification is related to sensitivity, adherence, or safety, then my discussion on the value of these properties will cast light on the value of justification. If justification is conferred by one of the truth-conducive properties discussed above, such as being based on good evidence or being the product of a reliable process, then the value of justification comes from the fact that it increases the chances that one will match the proposition. If justification confers only non-pragmatic value, then like the deontological properties, it would not be the appropriate subject for this sort of investigation.

---

[17] Roush, for instance, thinks that the value of truth tracking explains the value of knowledge.

The most interesting category of alternative or additional desiderata might be properties of systems of beliefs, such as coherence and completeness, and properties of processes for acquiring, maintaining, and discarding beliefs, such as speed and efficiency. One might think that the value we obtain from such desiderata does not reduce to the value obtained from having matching attitudes, or a high probability of matching attitudes, about particular propositions. Perhaps our chances of goal completion will be affected independently by the completeness of our body of beliefs, for instance, or by the efficiency with which we acquire relevant new true beliefs and discard false ones. Efficiency might have independent value if factors other than whether we match particular propositions play an important role in determining chances of goal completion. For instance, perhaps an agent's chances of goal completion would be greatly helped if she had a matching attitude about *p*. But to obtain a matching attitude about *p*, she would need to expend a great deal of energy and time, and these expenditures would seriously compromise her chances of goal completion. So, before investigating the chances of completing a particular goal by assessing the probability that one will match a set of propositions, it might be necessary to check whether the value of matching the propositions in that set would be outweighed or counteracted by other factors, such as expenditures of time or energy.

However, the pragmatic value conferred by some of these properties of systems of beliefs and processes that produce beliefs can still be reduced to the value of matching a proposition. The value of the speed at which we acquire and revise our beliefs seems to be accounted for in the probability that we will match *p* at the relevant times. If one's belief production process is slow, the probability that one will match in a given case is lower. Similarly, the value of completeness, in my view, derives from the fact that it is valuable to match some set of propositions. It is valuable to have a more complete body of beliefs because (in a simple case) if

there are three propositions that it would be valuable to match, it is better to match all three than just two. As far as coherence, I think that it is possible to incorporate anything that coherence might do to increase one's chances of goal completion into the assessment of which propositions and sets of propositions it would be valuable to match. Before one investigates whether we match or are likely to match a proposition or set of propositions, one investigates if it would be advantageous to match that proposition or set of propositions, and if it would not be advantageous to match a certain incoherent set of beliefs, this should be uncovered at that first step.

## 2.8 WHAT DETERMINES SENSITIVITY, ADHERENCE, AND THE PROBABILITY OF MATCHING? THE RELATIONSHIP BETWEEN CAUSES AND TRUTH TRACKING

The probability that an agent's attitude about a proposition will match the truth value of a proposition, as I have discussed, is determined by the interdefined terms Pr(p), Pr(Bp), Pr(~Bp|~p) (sensitivity), Pr(Bp|p) (adherence), Pr(p|Bp) (safety), and Pr(~p|~Bp) (safety counterpart). The causes of our moral attitudes determine Pr(Bp). Relations between these causes and the facts—whether it is the case that *p*—affect the agent's degree of sensitivity, adherence, safety, and safety counterpart.

As Roush notes, in practice humans only have truth-tracking beliefs because there are certain processes that exist between the fact of the matter and a belief about the fact of the matter. Roush writes, "it happens to be a contingent fact about human beings that we can't fulfill the tracking conditions without intermediaries: causal processes, one event indicating another,

one trait correlated with another, our eyes, our brains, having dispositions to respond differentially, testimony of witnesses, and so forth" (Roush 2010, 269). In my view, all the intermediaries that Roush mentions, as well as things like inferences and reasoning processes, are causal processes linking facts and beliefs. I should note, though, that in addition to direct causal connections between fact and belief, causal relations in which one thing causally influences both belief and fact of the matter (if there are relations of this sort in the moral domain) would affect probability of matching as well. In sum, the probability of matching is determined by causal relations in the world. Consequently, investigating the causes of our beliefs can potentially help us obtain information about the probability that our moral beliefs match.

To assess whether an agent has a high degree of adherence with regard to *p*, we should see whether we can obtain evidence about how much her attitude is influenced by factors that are not connected in the right way to the fact of the matter—e.g. factors that are morally irrelevant—or evidence that the causes of our beliefs made us unlikely to obtain the belief we have or likely to lose our belief in the future. Much of the literature on the debunking potential of evolution and other causes of our moral beliefs has focused on a problem that can be understood in terms of the failure of adherence without recognizing that adherence is the property in question. People point out that it could have easily been the case that the belief was not held (which is only noteworthy if there is no reason to think that the fact of the matter would have changed—people usually do not state this accompanying premise explicitly). To assess whether an agent has a high degree of sensitivity with regard to *p*, we will be looking for evidence about how heavily determined the belief is by potentially irrelevant influences or how inflexible one's attitude would be in the face of many variations in one's circumstances. If irrelevant influences seem to determine the belief or if the belief is unusually incorrigible, it is likely that the belief would be maintained even if

the proposition were false. In the next two chapters, I examine several psychological and evolutionary causes of our moral beliefs and I characterize three debunking arguments that can employ information about the causes of moral beliefs to produce conclusions about the probability that we match moral propositions relevant for achieving moral goals.

# 3.0 THE PRIORITY OF PROXIMAL CAUSES AND THREE PROXIMAL DEBUNKING ARGUMENTS

In this chapter, I will argue that we should start our investigation into what the causes of moral beliefs tell us about their epistemic status by looking at proximal rather than distal causes. I will begin by providing a general argument for the priority of proximal causes. Then I will turn to the three debunking arguments, illustrating each with a particular case. The first argument I discuss is what I call the dysfunctional component argument, which I will apply to the case of moral beliefs about helping that hinge on the activation of sympathy. The second is the argument from incompatible variability, which I illustrate with cases of moral political beliefs, punishment judgments subject to influence by emotion, and beliefs influenced by disgust. The third is the argument from inappropriately improbable belief, which I also illustrate with the case of beliefs influenced by disgust. In each of the cases I discuss in this chapter, we can reach conclusions about the epistemic status of moral beliefs influenced by proximal causes without examining the distal causes of moral beliefs.

## 3.1 PROXIMAL CAUSES AS A SUPERIOR SOURCE OF EVIDENCE ABOUT EPISTEMIC STATUS

I argue that the sort of information we have about distal causes of moral beliefs is

relevant for assessing the epistemic status of moral beliefs only if we cannot determine whether the proximal process underlying these beliefs is reliable just by looking at the properties of that proximal process. The influence of evolution, a source of more distal causes, will be relevant in some cases, but my thesis implies that any investigation into the epistemic status of moral beliefs given their causes should start with proximal causes.

The philosophical literature on the causes of moral beliefs can be divided into two broad categories: work dealing with more distal causes of human morality, especially evolution (whether biological or cultural), and those dealing with more proximal influences on moral belief, such as emotions. There has been much discussion recently about what the evolutionary origins of the psychological processes underlying moral beliefs imply for the epistemic status of moral beliefs, especially in response to Richard Joyce's (2001, 2006) and Sharon Street's (2006, 2008) work (e.g. Machery and Mallon 2010; Copp 2008; Enoch 2010; Schafer 2010; Kahane 2011; Brosnan 2011; Fraser 2014; Griffiths and Wilkins n.d.). Street has said that evolution does not pose a unique threat to moral beliefs, and that the same sort of problem that evolution raises could be posed by "any kind of causal influence on the content of our evaluative judgments" (Street 2006, 155). Despite this, evolution has attracted disproportionate attention in the literature. There has recently been some discussion, though, about what more proximal causes (such as emotions) of moral beliefs signify for the epistemic status of moral beliefs (e.g. Singer 2005; Greene 2008; Sinnott Armstrong 2008; Berker 2009; Joyce 2008, Kahane and Shackel

2010; Kelly 2011; Kahane 2012).[18] There is also a broader literature into which the topic of the causes of moral beliefs falls: literature on how causes affect epistemic status of beliefs generally (e.g. Cohen 2000, Sher 2001, Leiter 2004, White 2010, Vavova forthcoming, Joyce 2013, forthcoming, Sinnott-Armstrong 2005, Shafer-Landau 2012, Bedke 2009, ms, Setiya 2013). There are parallel discussions about the evolutionary or other distal origins of other types of beliefs, such as mathematical and logical beliefs (Schechter 2013, Clark-Doane 2012).

## 3.2 THE PRIORITY OF PROXIMAL CAUSES

My thesis says that we should give priority to proximal causes in our investigation of the epistemic status of moral beliefs. In this section I present a general argument for this thesis. As I argued in the previous chapter, the sort of epistemic status that should concern us is the probability that our epistemic attitudes about moral propositions match the truth value of those propositions. Beliefs are the product of many sorts of causes. Most directly, beliefs are the product of psychological processes. These psychological processes are produced by developmental processes. Developmental processes are themselves the product of evolutionary processes. Thus, beliefs are the product of a proximal psychological process, which is produced by (or is part of) a developmental process, which is itself the product of (or is part of) the more

[18] Andreas L. Mogenson also has a recent paper, "Evolutionary debunking arguments and the proximate/ultimate distinction," on the relevance of proximate and ultimate causes for debunking, but it addresses a question orthogonal to my topic here.

distal evolutionary processes.[19] By a distal cause of a belief, then, I refer here specifically to a *higher-order* process that produces the proximal process that produces the belief. Evolution is a distal cause of our beliefs in this sense because it produces the psychological processes within individuals that produce beliefs. The available information about the evolutionary process leading to our moral psychology that could shed light on the epistemic status of moral beliefs is either information about the probability that we would obtain the moral faculties that we have or ones that are similar, or information about the probability that evolution would produce reliable moral faculties.[20]

By contrast, examining properties of the moral beliefs themselves and of the proximal processes that produce them can provide excellent evidence about whether the moral beliefs and causal processes that we actually have possess the features that they would need to have if they were, respectively, true and truth tracking.

If we can ascertain whether the proximal process tracks the truth on the basis of features of that process, looking at the distal process that produced that proximal process provides us with no additional information about the epistemic status of our beliefs. By contrast, if we ascertain the probability that the distal process would produce a truth tracking proximal process, we can still gain additional information about the epistemic status of our beliefs by looking at the features of the actual proximal process that the distal process ended up producing. For instance, even if we determine that the distal process has a low probability of producing truth tracking

---

[19] This way of thinking of about the relationship between causal processes comes from C. H. Waddington (e.g. 1959).

[20] I return to the difference between these two kinds of information later in the chapter.

proximal processes, our proximal process may nonetheless, against the odds, be one that tracks the truth. In such a case, we can potentially gain additional information by investigating the proximal process directly. The only sort of situation in which information about the distal process alone could be sufficient to permit conclusions about the epistemic status of moral beliefs would be one where the distal process had a probability of 1 or 0 of producing proximal processes of a given degree of reliability. The evolutionary origins of the faculties that produce moral beliefs can provide us with *some* information about whether our moral beliefs track the truth, but there is an asymmetry in how much they can tell us compared with how much proximal processes can tell us. I propose that in principle information about proximal causes is sufficient for ascertaining whether our moral beliefs track the truth, but that information about distal causes is not sufficient for the same purpose.

If we could determine just by looking at features of our moral psychological processes that we track moral facts, we would not need to investigate how probable it was that evolution would produce truth tracking moral psychological processes. A parallel point holds if we can determine on the basis of features of our moral psychological process that we do not track the truth of moral facts. In such a case, we need not investigate the probability that a truth tracking psychological process would evolve. I am skeptical that we can ever conclude with confidence on the basis of information about the features of moral psychological processes that we track the truth of moral facts, but, as I show with the cases of disgust and sympathy, it is sometimes possible to conclude with confidence that we do not track the truth of moral facts without looking at distal processes.

In cases where we cannot conclusively determine the reliability or unreliability of our proximal processes on the basis of information about their features alone, we can use information

about the evolutionary origins of these processes to inform our assessment of the epistemic status of moral beliefs. For instance, if we cannot gather enough information about whether our current proximal process tracks the truth because there is conflicting evidence or we have an inadequate understanding of the proximal process, the probability that evolution would have produced a process that tracked the facts in question can provide additional information. Alternately, if we were interested in the epistemic status of moral beliefs across multiple generations, and the relevant psychological processes would be subject to continued influence from evolution, information about evolution could improve our estimation of whether those psychological processes will track the truth in future generations. These are the situations to which the evolutionary debunking arguments advanced by Street, Joyce, and others are useful—situations where we cannot already be confident about whether proximal processes track truth just by looking at the features of those processes. But if we can determine with confidence the reliability of proximal processes on the basis of the features of those processes, we need not look at distal causes in order to draw conclusions about truth tracking. Thus, when we set out to ascertain whether our moral beliefs track the truth, our first stop should be the empirical work on proximal psychological processes rather than work on the evolution of those proximal processes.

This point generalizes to beliefs outside morality. That is, information about proximal processes is more valuable than information about distal processes even when we are assessing the epistemic status of non-moral beliefs. As illustration of the generality of the priority of proximal causes, consider a case from outside the domain of morality of beliefs produced by proximal processes that are themselves produced by distal processes of varying degrees of

reliability.[21] Suppose there are two people, Juana and Marta, and we want to evaluate the epistemic status of their beliefs about certain mathematical propositions. Suppose we are confident that each acquired reliable mathematical reasoning processes early in life that regularly produce true beliefs about some set of mathematical propositions. We are confident that each will continue to hold these mathematical reasoning processes for the rest of her life. If we were only interested in these two people and their beliefs about the relevant set of mathematical propositions, and we could be confident that their mathematical reasoning processes were reliable, then we could be confident that each is tracking the truth of these mathematical propositions, regardless of where their reasoning processes came from.

However, if we were unsure about whether each person possessed a reliable mathematical reasoning process, we could obtain some information about whether they track the truth by looking at the distal process by which they obtained their mathematical reasoning processes. Suppose the distal process by which Juana obtained her mathematical reasoning process was reading a reputable mathematics textbook, and the process by which Marta obtained her mathematical reasoning process was listening to a local mystic who usually disseminates mathematical falsehoods and ineffective tricks. Looking at those distal processes provides us with information about the probability that each woman acquired a reliable mathematical reasoning process—high in Juana's case, low in Marta's case, Of course, the information we obtain by this route misleads us here, since in fact Marta against the odds acquired a reliable proximal process and does track the truth of the relevant propositions. We would be better off with good information about the proximal reasoning processes that Juana and Marta actually

[21] This case is a modification of Goldman's Humperdink case (1986, 51-52).

possess.

However, suppose we also have an interest in the epistemic status of the mathematical beliefs of others in the same society, in particular, in the beliefs of the young daughters of Juana and Martina, who have not yet acquired their mathematical reasoning processes. Even if we know that Juana and Marta have reliable mathematical reasoning processes, it would be helpful for the purpose of gauging the epistemic status of the beliefs of others to know that Juana obtained her mathematical reasoning processes by textbook and Marta by mystic. If the next generation obtains its mathematical reasoning abilities from the same sources, it is very likely that some of them will end up with unreliable mathematical reasoning processes and will consequently be likely to believe mathematical falsehoods and lack beliefs in mathematical truths. Thus, distal processes can add information for our assessment of the status of beliefs in cases where we cannot assess the reliability of proximal processes of interest, whether due to limited information about the proximal process or because the distal process continues to exert influence on the proximal processes that will produce the beliefs of interest.

Thus, distal processes can supply useful information for our assessment of the status of beliefs in cases where we are unsure about the reliability of proximal processes. In other cases, distal processes do not provide information for assessing epistemic status beyond what can be gained from information about proximal processes.

I have argued that any investigation into the epistemic status of moral beliefs given their causes should start with proximal causes. In cases where we are confident in our estimation of the degree to which a proximal process tracks the truth, we need not look at the distal origins of those proximal causes in order to draw conclusions about the epistemic status of the moral beliefs that they influence.

## 3.3 THREE DEBUNKING ARGUMENTS

In this second part of the chapter I introduce three debunking arguments that can be constructed using information about the causes of moral beliefs and that minimize normative assumptions. Each argument is of potentially narrow scope, applying to just a subset of moral beliefs. I apply each argument to one or two cases of types of moral beliefs. Each of these arguments also employs etiological information only about proximal rather than distal causes, serving as examples to demonstrate that information about properties of a proximal cause can be sufficient to conclude that moral beliefs influenced by that cause have a poor epistemic status. The arguments I introduce are the following: (1) the malfunctioning component argument, which I apply to the case of moral beliefs that hinge on sympathy, (2) the inconsistent variability argument, which comes in two forms (a) attitude hypervariability, which I apply to the case of (i) punishment judgments subject to variation due to the influence of emotions, and the case of (ii) political beliefs that change due to factors that vary with age, and (b) attitude hypovariability, which I apply to the case of moral judgments that have been influenced by disgust (incorrigibility), and lastly (3) the Inappropriately Improbable Belief Argument, which should lead one to distrust judgments influenced by disgust as a result of disgustingness's property of transmissibility.

### 3.3.1 The Malfunctioning Component Argument: Sympathy

The first argument I propose identifies a component in the moral belief production process for which we have independent methods for assessing reliability, and shows that the component is malfunctioning. I am not claiming that this is any sort of evolved or designed function, nor that

the function is ever successfully performed; I am claiming that we can sometimes identify the function that the component would need to perform if it were to contribute to the process tracking the truth. The idea is that we can sometimes infer that moral judgments are not truth tracking because the moral judgment hinges on a non-normative judgment that we know does not track whatever it *would need to be tracking* for the moral judgment to track the truth.

In the case of moral beliefs that rely on sympathy, the targeted component is malfunctioning sympathy activation. If sympathy were to contribute to truth tracking moral beliefs, I take it that it would do so by picking out entities that are suffering. I show that at this task sympathy produces both false positives and false negatives at a significant rate, and conclude that moral beliefs that hinge on the reliable activation of sympathy are neither sensitive nor adherent and we should reduce our confidence in such beliefs.

This argument might appear to bear a resemblance to what Nichols (2014) calls a "process debunking argument." Nichols (2014) characterizes the process debunking argument as an argument in which one concludes that a belief is unjustified by showing that a belief is the product of a process that is known to be epistemically defective (e.g., distorting or unreliable), such as dreaming or wishful thinking (2). However, the malfunctioning component argument is a specific method for showing that the process leading to moral beliefs is epistemically defective (e.g. that it is unreliable or that it does not reliably produce matching beliefs about a set of propositions); the method is to show that a component of the process is performing whatever function it would need to perform for the process as a whole to be reliable.

The psychological cause I will use to illustrate this argument is the emotional process of feeling bad for someone else, or what Maibom calls "affectively infused concern for the other's

welfare" (2012, 62).[22] I use the term "sympathy" to refer to this process, but much research on "compassion" appears to be concerned with the same phenomenon (for a review of the compassion literature, see Goetz 2010).[23] Frequently when we judge that we or someone else should aid a person or other entity, we are led to do this by the process involving sympathy. Certainly there are also cases where we judge that a person should be given help without experiencing sympathy—for instance, policymakers may decide they should allocate resources to a group they know to be needy without feeling sympathy for the members of the group. The moral judgments that I claim are not truth tracking are those that are the product of a process in which sympathy activation plays a substantial role: situations where if one feels sympathy, one is much more likely to judge that an entity should be helped, and if one does not feel sympathy, one is much less likely to make the judgment that the entity should be helped.

I take it that if sympathy activation were ever to help us make matching moral judgments about propositions related to whether someone ought to help another entity, it would do so by accurately picking out entities that are suffering or that are otherwise are in a situation that could conceivably be improved. Consequently, we should be concerned about how reliably sympathy picks out those entities that are in fact in pain. There is reason for concern in two directions:

---

[22] Similarly, Pfattheicher et al. characterize sympathy as "concern for suffering others" (3). This has also been called "empathic concern," Batson studied something like this under the name of empathy, and it may the same thing that Prinz calls "concern."

[23] I focus on sympathy/compassion rather than on related processes like empathy or emotional contagion or distress, because I take it that these latter psychological processes have less a connection to the judgment that an entity should be helped.

reliance on sympathy produces both false positives and false negatives. It sometimes leads us to classify entities as in pain when we have independent reason to believe that they are not in pain, and it sometimes leads us to classify entities as not in pain when we have independent reasons to believe that they *are* in fact in pain.

On the false negatives side this mechanism of pain attribution appears less likely to engage when it comes to outgroup members. People often empathize less with outgroup members experiencing pain than ingroup members. For instance, Trawalter and colleagues (2012) show that white subjects tend to attribute lower levels of pain to black people (see also Cikara and Fiske 2011).

Another case of false negatives comes from the well-known human disposition to feel sympathy for non-human animals with certain features and not for non-human animals that lack those features but that we have equally good reason to think experience comparable levels pain.

On the false positives side, consider recent work on robots. Rosenthal-von der Putten and colleagues (2012) found that subjects who watched videos of robots being treated violently exhibited higher physiological arousal, expressed empathy for the robot, and reported more negative emotional reactions than subjects who watched videos of robots being treated in a friendly way. Rosenthal-von der Putten and colleagues also found similar patterns of brain activation when subjects watched videos of humans and robots being treated affectionately (though admittedly somewhat higher brain activation when subjects viewed humans being treated violently than when they viewed robots being treated violently). Anecdotal evidence about the use of companion robots like "Paro," a seal-like robot, and "Keepo," a robot that looks like an Easter peep, also suggests a tendency to respond to entities that exhibit certain features, such as cuteness and responsiveness, as if these were entities capable of suffering. Anecdotal

reports indicate that soldiers frequently become "fond of" the robots they use in battle and that some soldiers have actually risked their own lives to rescue robots—presumably not just because the robot was a valuable piece of machinery (Singer 2009). Also, work by Bartneck and Hu (2008) found that, although when instructed to do so, subjects destroyed robots with which they had interacted, the subjects expressed hesitance and regret, saying things like "I didn't like to kill the poor boy," "The robot is innocent," and "This is inhumane!" (426).[24] (For further discussion of this case and the features that lead subjects to attribute pain to robots, see Fiala, Arico, and Nichols 2014.)

If it is a sympathy-dependent process that leads us to think we should, for instance, avoid stepping on insects, prohibit abortion, or keep patients in a persistent vegetative state alive, the fact that this emotional pain-detection process so easily produces false positives may be reason to reduce confidence in our beliefs and to make our judgments about whether the entities in question experience pain using a method that does not hinge on sympathy. Likewise, if our judgments as policy makers or individuals about how to distribute resources or whom to assist are heavily dependent on the activation of sympathy as we consider each case, then in contexts where we are not feeling sympathetic we should decrease confidence in the conclusion that we need not aid the entity under consideration.

This is not to say that it would be better to ignore sympathy completely when detecting entities in pain in the world—it may yet serve as a heuristic in contexts where we lack alternative

[24] Of course, it is not inconceivable that a sufficiently complex robot could experience pain, and in particular pain in whatever sense is morally relevant. However, in the case of these simple robots, I think we have no reason to think that they experience pain.

methods for ascertaining whether an entity is suffering. Rather, the suggestion is that we should reduce confidence in beliefs involving pain attributions that are substantially determined by whether we experience sympathy, at least until we have identified the features that characterize the contexts where our sympathy-dependent pain attribution judgments go astray. In addition, it may be possible to refine or improve our sympathy response somehow in the future.

Bloom (2013) and Prinz (2011) have also recently argued against the reliability of empathy in moral contexts; some of their arguments seem to apply just as well to what I am calling to sympathy.[25] However, their arguments are different from mine in ways worth distinguishing.

Bloom offers several reasons to think that empathy is unreliable. Several of these reasons stem from the idea that "The power of this faculty [empathy] has something to do with its ability to bring our moral concern into a laser pointer of focused attention" (2013, n.p.). He highlights the Baby Jessica case and similar cases where public attention focuses on one person who is suffering and little attention is paid to many others suffering as much or more. Research suggests a number of reasons this happens, such as the identifiable victim effect. I interpret this problem for the status of our moral judgments in the following way: if there are a number of people experiencing an equal level of suffering, and our judgments about who we or others should help

---

[25] Sam Harris (2014) also has a related argument, against moral decisions guided solely by empathy. He says, "To be moved to action merely by empathy is to lurch blindly toward who knows what." His argument seems to be that judging that some act is wrong solely because of the activation of empathy is problematic because one fails to weigh the badness of the act against other considerations.

are substantially determined by empathy, we cannot be tracking the truth of whether any particular person ought to be helped, or perhaps how much each should be helped, because our judgments about these questions depend on our level of empathy for that person, which depends on whether that person happens to be an identifiable victim or has attracted the attention of the public. Bloom's argument is slightly more complicated than the one I introduce above, where I argued that sympathy is not dependable for identifying presence of suffering, which is a requirement for making matching judgments about whether another person ought to be helped. Bloom's argument has to do with how assistance should be allocated when there are multiple entities that are suffering. Our judgments on this question are affected by identifiability of the victim and the level of attention they attract, and these factors do not track levels of suffering; in addition, Bloom's argument needs to assume that identifiability of the victim and the level of attention they attract do not themselves make a difference to whether one ought to help the victim and do not track something about the victim that affects whether one ought to help him or her. Presumably many people will be willing to accept those assumptions, but they are worth noting.

Bloom also argues that if we feel worse about 20,000 people dying than 2000, it is due our rational capacity to compare the numbers, not empathy. Empathy does not respond differentially to these scenarios in a way that reflects the numbers if each life were worth the same. I interpret the problem this argument poses in the following way: we must not be tracking how much assistance should be given to a suffering group of 20,000 because an empathy-driven process would involve as much empathy in response to that group as to a group of 2,000, and if this process determines our judgment about how much help each group should receive, we will treat them as equivalent. Yet, we can assume that a larger group of suffering entities warrants

more assistance than a smaller group, all other things being equal, and so our process fails to detect a morally relevant difference in these two cases. Again the normative assumption that has to be made here is one that many will be willing to accept, but it is worth noting. Prinz makes a similar argument.

Bloom also sketches a number of other problems with empathy that I think require more substantial normative assumptions to flesh out. He observes that empathy may make us more likely to focus on immediate problems than more important but less salient problems—this is a suggestion that requires a standard for assessing the moral saliency of problems. Feeling empathy for victims, he says, may also motivate subjects to punish perpetrators even if the long-term consequences will be worse if the perpetrators are punished. For this argument, we need a standard for assessing the badness of consequences, which is a much more substantial demand than the assumption that more suffering people should be given more aid than fewer suffering people (or even more minimally that the number of suffering entities affects how much aid should be provided), that the identifiability of a suffering entity and whether he or she has attracted public attention do not track differences in how much help that entity should receive, or, as I assumed, that if sympathy were to contribute to truth tracking moral processes it would do so by reliably picking out entities that are suffering.

Prinz also identifies several problems with empathy: it is not substantially involved in "thinking about inequality, the environment, or cases where large groups of people are threatened," yet these are "core aspects of moral reasoning"; "empathy is easily manipulated, leading us to give preferential treatment to those who don't deserve it"; and "empathy is biased: it increases when those in need are salient, similar to ourselves, and close by." This first argument involves the idea that there are moral considerations related to inequality and the

environment that are worth attending to (something I take to be a substantive normative assumption) but that empathy is not good at tracking; the second requires a normative account of desert, and the third requires plausibly disputable assumptions about the moral irrelevance of saliency, similarity to oneself, and physical distance.

The argument that I have presented says that sympathy, on which some moral judgments depends, is unreliable at picking out what it would need to pick out if it were to help such moral beliefs track the truth. Consequently, we should decrease confidence in moral beliefs that are the product of a process substantially relying on sympathy activation. The malfunctioning component argument can be applied to a number of other contributors to moral beliefs, such as judgments about actors' intentions, the causal structure of a series of events, and judgments about expected consequences of a given action. The accuracy of judgments about intentions, the causal structure of the world, and expected consequences of actions can be investigated independently of normative questions. If we can show that important contributors to moral beliefs are unreliable or subject to errors that we can identify independently, we can conclude that the moral beliefs themselves are not tracking the truth.

### 3.3.2 The Argument from Inconsistent Variability

The second debunking argument I propose operates by establishing that a cause of one's attitudes confers on them properties that are inconsistent with plausible general features of moral propositions. In particular, the general feature of moral propositions that the argument from inconsistent variability appeals to is that the truth value of the moral proposition varies in some particular way over a time period of interest.

This debunking argument emerges from a strategy for assessing one's epistemic position

with respect to a proposition under a condition of limited information, in which one has little good first order evidence for or against the proposition. The attraction of this type of debunking argument is that one is able to construct it without knowing the truth value of the proposition of interest at any point.

The strategy consists of comparing the level of variability in the truth value of a proposition of interest with the level of variability in one's attitudes about that proposition, over some time period of interest.[26] If the level of variability of the truth value of a proposition is dissimilar to the level of variability in a person's attitudes about that proposition, then some proportion of the person's attitudes about the proposition over time are in error. Evidence of inconsistent levels of variability is evidence that one does not track the truth of the proposition. In the moral domain there are situations where we know that the truth value of the proposition in question does not vary across the times during which we consider it, even though we may not know whether the proposition is true or false at any of those times, so this debunking argument has particular value in the moral domain.

I will begin my introduction of this argument by demonstrating the epistemic significance of inconsistent levels of variability and discussing how we can identify inconsistent levels of variability in moral attitudes and the truth value of propositions. Then I will apply these general

---

[26] I will assume that the truth value of propositions can change over time. For those who take the position that the truth value of propositions does not change over time, my argument may be translated into different terms: the comparison will be between the diversity in the timeless truth value of the propositions expressed by uses of a given sentence over the time period of interest and the variability in one's attitudes about the propositions expressed by uses of the sentence.

points to moral beliefs, discussing two illustrations of attitude hypervariability, one form of the inconsistent variability argument, and then one illustration of attitude hypovariability, the second form of the inconsistent variability argument. The cases that illustrate attitude hypervariability are: (i) moral political beliefs that vary with age and (ii) judgments about punishment, which are vulnerable to influence by extraneous emotions. The case that illustrates attitude hypovariability is the case of moral beliefs that are the product of a process that includes the influence of disgust. I conclude that none of these classes of belief are truth tracking and that we should substantially reduce our confidence in these types of beliefs.

### 3.3.3 The Epistemic Significance of an Inconsistency in Levels of Variability

An argument from inconsistent variability draws epistemic conclusions from an inconsistency in the level of variability of the truth value of a proposition and the level of variability in whether a subject has or lacks a belief in the proposition. S's attitudes about *p* exhibit a high level of variability if S's attitude about *p* frequently changes—e.g. from belief that *p* to belief that ~*p*, or from no belief about *p* to belief that *p*—over some duration of interest. The truth value of *p* exhibits high variability if the truth value of *p* frequently changes over the duration of interest. An inconsistency in levels of variability could happen in two ways:

*Attitude hypervariability*: S's attitudes are more variable than the truth value of *p.*

*Attitude hypovariability*: S's attitudes are less variable than the truth value of *p*.

When we investigate the levels of variability in attitudes and the truth value of *p*, there are three possible outcomes: attitude hypervariability, attitude hypovariability, or levels of variability that are consistent with S routinely matching *p*. If S's attitudes about *p* are to match

the truth value of *p* most of the time, or, in other words, for S to tend to believe *p* when *p* is true and lack belief in *p* when *p* is false, it must be the case that S's attitudes vary at approximately the rate at which the truth value of the proposition varies over the duration of interest. This is a necessary condition for S's matching *p* most of the time. It is not sufficient, though: attitudes could change at the same rate as the rate at which the truth value of the proposition changes but nonetheless fail to match (e.g. if S always lacked belief in *p* when *p* and believed *p* when not *p*).[27] So, if we find that S's attitudes change about as much as we think the truth value of *p* changes, this is consistent with S's attitudes tending to match the truth value of *p* over the duration of interest, but we cannot conclude much more.

Discovery of an inconsistency in levels of variability, by contrast, can support more significant epistemic conclusions. Each of the two types of inconsistency supports different conclusions. In an extreme version of attitude hypervariability, the truth value of *p* is mostly static; in such a case, *p* will be either usually true or usually false. Now, if over the time period of interest *p* is mostly true, and S's attitudes about *p* vary substantially, S is frequently failing to believe *p* when *p* is the case: there is a failure of positive matching. If this sort of pattern is probable, it cannot be that S is likely to believe *p* given *p*, so S does not have a high degree of adherence with regard to *p*. If over the time period of interest *p* is mostly false, and S's attitude

---

[27] It could also be that over the time period we compare variability levels, there is the same overall level of variability, but there is variation in variability levels over that time period. In this case, overall levels of variability are consistent with matching but there could nonetheless be little matching. So, similar overall levels of variability do not provide great evidence for matching.

about *p* changes a great deal, S will frequently believe *p* in cases where *p* is false—there is a failure of negative matching. If this sort of pattern is probable, it cannot be the case that S is likely to not believe *p* when *p* is not the case, and so S does not have a high degree of sensitivity with regard to *p*. In short, attitude hypervariability implies that S is insensitive or nonadherent with regard to *p*. To reach a *specific* conclusion about sensitivity or adherence—either the conclusion that S is not sensitive or the conclusion that S is not adherent—requires a premise about whether *p* tends to be true or whether it tends to be false. Without that premise, in a situation of limited information, we can only conclude that either *p* is usually false and S is not sensitive, or *p* is usually true and S is not adherent.[28] Regardless, from evidence of a significant difference in levels of variability we can infer that S's epistemic situation is poor.

For instance, suppose S wants to know whether the temperature outside is between 50-60 degrees, and S forms beliefs on this question solely via a thermometer outside. S could go outside and check what the thermometer says, but from inside S cannot see what the thermometer reads. What S can see from inside, though, is that the mercury in the thermometer is all over the place—it varies wildly over the course of, say, ten minutes. Even though S does not have any sense of the temperature outside or of what the thermometer reads at any particular time, if her background knowledge of how the weather operates in the area indicates that the temperature outside always varies very little over the course of ten minutes, she can conclude that her process of determining the temperature produces attitude hypervariability. Consequently,

[28] Given that the variation in S's attitudes is substantial, a reliabilist could also infer that the process leading to S's attitudes about this proposition is not reliable, in the sense that true beliefs are not a sufficiently high percentage of the attitudes it produces.

S can conclude that she is in a poor epistemic position with regard to the proposition that it is 50-60 degrees outside: she is insensitive or nonadherent with regard to that proposition. Either it is false that she is likely to not believe *p* given that *p* is false; or it is false that she is likely to believe *p* given that *p* is true.

The second type of inconsistency between levels of variability, attitude hypovariability, occurs when S's attitude changes less frequently than the truth value of *p*. At the extremes, S either tends to believe *p* or S tends to not believe *p*. If over the time period of interest, S tends to believe *p*, S will frequently believe *p* in those cases where *p* is false (failure of negative matching). Consequently, supposing this pattern of attitudes is probable, S must not have a high degree of sensitivity. On the other hand, if over the time period of interest, S tends to not believe *p*, S will frequently be failing to believe *p* in those cases where *p* is true (failure of positive matching). Consequently, S must not have a high degree of adherence. Thus, if S could establish that the truth value of *p* changes a great deal but that she tends to stubbornly hold a belief in *p* or stubbornly lack a belief in *p*, she could draw a conclusion about sensitivity in the first case or adherence in the second case.[29]

Thus, as in the case of attitude hypervariability, discovery of a difference in levels of variability implies that S's epistemic situation is poor. For instance, suppose that S looks at an

[29] Furthermore, from attitude hypovariability, S can also infer that her process for forming beliefs about *p* is unreliable in the following sense: whether she tends to believe *p* or not to believe *p*, if the truth value of *p* varies a great deal, S's process will not produce a sufficiently high proportion of true beliefs out of the set of attitudes her process produces on the subject matter.

ordinary clock that reads "1:00."[30] The proposition in question is that it is 1:00. S is interested in the truth value of this proposition at the moment she looks at the clock. S knows that the truth value of *p* changes over the course of the day, even if S does not know the truth value of *p* at any given moment. If S discovers that this clock shows no sign of changing the time it reads over some long duration, such as an hour, and if she bases her beliefs about the time on a reading of the clock, she may conclude that there is an inconsistency between her beliefs and the truth value of the proposition, in the form of attitude hypovariability. S's belief does not change frequently enough to be matching the fact of the matter. On the basis of this inconsistency, S can infer that she is in a poor epistemic situation with regard to the proposition. She is insensitive with regard to the proposition that the time is 1:00, because when it is false that the time is 1:00, she is nonetheless highly likely to believe that it is 1:00.[31]

Discovery of attitude hypervariability or attitude hypovariability indicates that we are erring—failing to match—in some proportion of the situations that concern us. But such a discovery has the potential to show not just that errors are being committed, which we accept as inevitable in most of our belief-forming processes, but also in some cases that errors are occurring at a significant enough rate to mean that we are in a poor epistemic position, failing to track the truth of the propositions that interest us. In an argument from inconsistent variability as

---

[30] Suppose the clock, when operating properly, adjusts its hands with discrete movements once per minute.

[31] Furthermore, S's process for forming attitudes on the matter of whether it is 1:00, or, even more broadly, on the matter of what time it is, is unreliable (regularly producing the false belief that it is 1:00).

I formulated it above, evidence of a significant inconsistency in variability between beliefs about p and the truth value of *p* can lead to conclusions about our epistemic status with regard to that proposition in terms of sensitivity and adherence.

### 3.3.4 How We Can Identify Inconsistent Levels of Variability

To construct an argument from inconsistent variability we need to identify the levels of variability in our attitudes toward *p* and in the truth value of *p*. We can often investigate empirically the variability of a person's attitudes about particular propositions or sets of propositions, including moral propositions. For example, we can assess the level of variability of a person's attitudes about *p* from observed variation in the possession of attitudes about *p* over some duration. We can use this information to estimate the level of variability that the person's attitudes will exhibit in other contexts or in the future. But we can also investigate the causal processes that produce the person's attitudes about *p* to discover how likely she is to believe *p* in various conditions she is likely to encounter. I give examples of each of these methods for assessing variability in a later section: in one case I argue that political attitudes are likely to vary over one's lifetime by using evidence about how people's attitudes have tended to vary in the past; in a second case, I show that individuals' attitudes about punishment are likely to vary by using information about the causes of these attitudes.

We can identify the level of variability in the truth value of a proposition in several ways. A direct way is to infer the variability in the truth value of a proposition over some time period from information about the truth value of the proposition at various times or factors that vary in accordance with the truth value of the proposition. With this sort of information we can gauge the proposition's level of variability, but if we have such information my proposed strategy of

assessing variability levels is unnecessary for evaluating the epistemic status of our beliefs—such information will permit conclusions directly about reliability and truth tracking. But this direct approach has limited utility in the moral domain, because part of the problem with assessing the epistemic status of moral beliefs is that there is substantial dispute over the truth value of moral propositions, as well as dispute over which causes or features of a situation are morally relevant and which belief-forming processes are reliable. For instance, if one were confident that it is wrong to push a large man off a bridge to stop a trolley, one could criticize a process that tends to lead us to the belief that pushing the man is right, and one could claim that people who believe this are failing to track the truth. However, there is significant dispute over whether pushing the man is right. Similarly, if one were confident that disgust is a distorting influence on moral judgments, one could conclude that moral beliefs heavily determined by disgust reactions are not the product of a reliable process. However, there is dispute about whether disgust is a good or bad guide to the immorality of actions (see, e.g., Kelly 2011, Kass 1997). It would be better if we could avoid these assumptions.

Where this direct approach is not feasible there are several ways in which the level of variability of a proposition can be assessed. For instance, one might have background knowledge that informs one's estimation of the variability of the truth value of the proposition, as in the examples discussed above, where we have background knowledge about the weather or about time. Or one may have reasons to think that the proposition has a certain level of variability in virtue of the domain to which it belongs. For instance, one may be confident that a logical or mathematical proposition is a matter of necessity so that its truth value does not vary over the period of interest even though one may not yet know its value or may lack a method for

ascertaining it. Lastly, one may know that the truth value of the proposition of interest is heavily influenced by some causal factor, and one knows how much that causal factor varies.

In the moral domain, I suggest that metaethical premises can supply the background assumptions that enable one to draw conclusions about the stability of the truth value of certain types of moral propositions over various time periods without making normative assumptions. Some have argued that certain types of propositions in the moral domain are a matter of conceptual or metaphysical necessity (e.g. Swinburne 2004). If one thinks this, one may be confident that the truth value of those types of propositions will not vary over whatever period is of interest. Alternately, one might think that the truth value of moral propositions is or is not affected by some factor, such as social agreement or human nature, and one has a sense of how much that factor varies in the time period of interest. For instance, as long as one knows that the relevant society has some view on the proposition in question, a contractarian need not know whether the society considers the proposition true or false to infer that the truth value of the proposition will not change over very short time periods. Of course, any claim about what determines the truth value of moral propositions is a significant claim to make. However, the key for my argument is that different views about the relevance of various factors overlap substantially in the level of stability in the truth value of moral propositions that they imply.[32] The realist might think that it is a mind-independent fact that one should not cause harm on a whim; the constructivist might think that it is true that one should not cause harm on a whim because our society agrees upon this; an egoist might think that one's long term self interest makes it false that one should not cause harm on a whim. None will think that the truth value of

[32] Some (though not all) subjectivist accounts constitute an exception.

this proposition will vary over the course of a day or a week. If a person alternates day to day between thinking that she should never cause harm on a whim and thinking that she may sometimes cause harm on a whim, proponents of each of these metaethical views should be able to agree that there is a mismatch between her attitudes and the truth value of the proposition, regardless of what that truth value is.

### 3.3.5 The Inconsistent Levels of Variability Argument Applied in the Moral Domain

#### 3.3.5.1 Overview

Here I will first show that in the moral domain an argument from attitude hypervariability can demonstrate that many people are in a poor epistemic situation with regard to certain types of moral propositions. This is because many people's attitudes about moral propositions of those types are likely to vary even though the truth value of the relevant propositions does not vary. In other words, in these cases people are likely to believe a moral proposition *p* at some times and to not believe *p* at other times, when the truth value of *p* has not changed. I will develop an argument from attitude hypervariability for the following two types of moral beliefs: (a) moral beliefs in the political domain, which are too likely to vary as one ages, and (b) moral judgments about punishment, which are too likely to vary depending on one's emotional state. In each case, the presence of attitude hypervariability should lead us to substantially decrease our confidence in certain moral beliefs in the political domain and beliefs about punishment. Then I will show that we can also develop an argument from hypovariability in the moral domain, using the case

of moral beliefs influenced by disgust, which exhibit a high degree of incorrigibility that makes them too invariant to be tracking the truth.

#### 3.3.5.2 Attitude Hypervariability: Case 1: Variability in attitudes on moral questions in politics over a lifetime

Here I will discuss variability in moral attitudes related to politics over a lifetime.[33] There is evidence that as people age they tend to become more conservative on a variety of political questions such as those related to harsh treatment of criminals, censorship, and pacifism (Hewitt et al. 1977; Truett 1993; Cornelis et al. 2013).[34] Ray (1985) found that older people scored higher on a scale for general conservatism that included such claims as "Patriotism and loyalty to one's country are more important than one's intellectual convictions and should have precedence over them," "Treason and murder should be punishable by death," and "We should have

[33] The question of the epistemic significance of variation in belief with age also arises in Colaço et al. (2014) in the context of intuitions about knowledge.

[34] See also Ojha and Sah (1990); Lupfer and Rosenberg (1983) conclude that an increase in conservatism with age is mediated by life events. There is also evidence that one's socioeconomic class may affect moral judgments (Côté et al. 2013), as may the subject's level of social dominance (Armstrong 2013), each of which are factors that are reasonably likely to change over a lifetime. There are other topics on which people's attitudes are unlikely to vary with age, and some topics on which people's attitudes are likely to change in accordance with changes in the population's attitude on the topic (Glenn 1974; Cutler and Kaufman 1975).

complete freedom of speech even for those who criticize the law" (527; see also Grant et al. 2001, Henningham 1996). The effect of interest here is not that the population as a whole becomes more conservative over time (a period effect), nor that each generation is more liberal than the last (a cohort effect); the point is that in each generation a given individual who is liberal in her youth is likely to be more conservative when she is older (a life-cycle effect).[35] Several longitudinal studies support this claim. Marwell et al. (1987) performed a longitudinal study on liberal activists from the 1960s that identified a lessening of liberal attitudes on questions of non-violence and civil disobedience over twenty years. Konty and Dunham (1997) found a shift toward conservatism in attitudes about the importance of law and order and U.S. power. Studies like these supply us with evidence that an individual's political attitudes are likely to vary over the course of her lifetime.

We should be able to agree that the truth value of at least some of these contested moral questions in the political domain, such as whether the death penalty is ever an appropriate

---

[35] Multiple studies conducted at different times over the last century (Eysenck 1975; Hewitt et al. 1977; Feather 1978; Truett 1993; Riemann et al. 1992) have established a correlation between age and conservatism. This supports the claim that older people are consistently more conservative than younger people, but the phenomenon could be explained either by individuals becoming more conservative or by a cohort effect in which younger generations are always more liberal than older generations but each individual's attitudes are constant. Some of differences in attitude undoubtedly *are* due to a cohort effect (see discussion in Sears 1981), but there is also evidence that greater conservatism in older people is in part due to variation in the attitudes of individuals.

punishment, does not vary as individuals in a population age. An individual whose belief on one of these political question changes over time does not think that her current view and her past or future view might both be correct.

So what should we conclude from a pattern showing that the liberal beliefs of a young person are likely to be replaced with conflicting more conservative beliefs? Take a young person, Eloise, who believes that torture should never be used, no matter the context, and even if it would save lives. Faced with evidence that she is likely to abandon this belief later in life, should she revise her degree of confidence? As an older person who believes, contrary to her earlier position, that torture may be used under some conditions, what should she do when she learns about this trend in attitude change? Should the older Eloise also reduce confidence in her belief? The truth value of the proposition that torture is never permissible does not vary over the course of her life, yet for some of her life, she believes the proposition and for some, she does not. Thus, whether it is true or false that torture is never permissible, she knows that over the course of her life, she is in a poor epistemic situation with regard to the proposition that torture is never permissible—she does not track the truth of this proposition over the course of her life, and she does not obtain her attitudes about this proposition via a reliable process. So, whether it is in her youth or when she is older that she learns about a trend toward conservatism, she should reduce her confidence in her belief about the proposition that that torture is never permissible.

Notice that we are able to reach this conclusion without taking a stance on whether torture is permissible, and the argument does not require a premise about particular causes of beliefs being irrelevant or distorting, nor processes being unreliable. In fact, we can infer from attitude hypervariability that her belief-forming process is unreliable or that there is some irrelevant factor influencing her attitudes. Thus, the argument from irrelevant factors avoids

normative assumptions that the arguments from unreliable processes and irrelevant factors rely on, but allows us to draw significant moral epistemological conclusions.

A natural response to the argument applied in the Eloise case is to say that Eloise need not reduce confidence in her belief *at one of these times* if she can show that she is in a better epistemic position at that time than at times when she holds the conflicting belief. To do this would be to isolate the epistemic problem. The process that leads Eloise to beliefs about torture over the course of her life may be unreliable, but perhaps the process Eloise employs to develop her belief on this question when she is older, for instance, constitutes a reliable subprocess. If she could show this, she could retain her level of confidence in her belief that torture is never permissible. But until this has been done, she should reduce confidence in her belief no matter her age.

In the case of moral-political questions on which one's views are likely to vary over the life cycle, I think we lack a clear-cut reason to prefer political beliefs held when younger or vice versa. One might be inclined to think that people are in a superior epistemic position when they are older, for various reasons. One could appeal to a general epistemic principle that says that across domains one tends to get more accurate over time—maybe because one gains greater wisdom or acquires more evidence. Against such a general principle, there is also evidence from outside the moral domain that older people exhibit various biases that make them more likely make incorrect judgments on certain tasks (Glisky 2007). There are also differences between the cognitive capacities of younger and older people, such as greater inflexibility and changes in personality and emotional responses as one ages (Levenson 2000), whose general epistemic upshot is unclear.

Some might argue that one would be justified in the face of attitude hypervariability in simply maintaining confidence in whatever belief one happens to have at the moment. There is a position like this in the literature on peer disagreement; it has been argued that one may justifiably maintain one's level of confidence in one's belief in the face of peer disagreement (e.g., van Inwagen 1996; see discussion in Kelly 2005). It may be the case that in the face of peer disagreement one is entitled to give special weight to one's own belief and may thereby be justified in maintaining one's level of confidence. In the case I am discussing, though, the conflicting beliefs are held by the same individual at different times. It goes too far to say that one has an entitlement to give special weight to the belief one has at a particular moment when one knows that one's own belief is highly likely to be different subsequently.

So, for any moral proposition in the political domain whose truth value will not vary within one's lifetime, if one is given the information that one's beliefs about that proposition are likely to vary, and if one lacks a way to isolate the unreliability, the argument from inconsistent variability says that one does not track the truth of that proposition. This means, among other things, that both young liberals and older ex-liberals should reduce confidence in their beliefs about certain propositions related to punishment, authority, patriotism, pacifism, and threats to law and order.

#### 3.3.5.3 Attitude Hypervariability: Case 2: Variability in attitudes about punishment influenced by externally induced emotions

A variety of externally induced emotions have been shown to influence moral beliefs (e.g., Tiedens and Linton 2001; Valdesolo and DeSteno 2006, 2008; Strohminger et al. 2011; Conway

and Gawronski 2013). Polman et al. (2012) found that externally induced anger, guilt, or envy can affect moral judgments, such as judgments about the acceptability of speeding, tax dodging, or keeping a stolen bike, or how much subjects think others should donate to charity. Inducing a feeling of disgust from a source other than the vignette that subjects are asked to consider (for instance, via hypnotism) can also have an effect on subjects' judgments about the vignette (Wheatley and Haidt 2005). Priming subjects to feel clean or pure can also influence moral judgments (Schnall et al. 2008a; Zhong et al. 2010; Helzer and Pizarro 2011).

Here I will focus on the influence of externally induced emotions on judgments about punishment. To establish attitude hypervariability using information about the causes of one's attitudes, one needs to show (1) that the process by which one forms a moral judgment produces different judgments depending on variation in some causal factor, (2) that that factor is likely to vary over the duration of interest, and (3) that the fact of the matter (if there is any) does not change over the duration of interest. In this case the three conditions are met in the following way: (1) varying externally induced emotions can produce variation in judgments about punishment, (2) emotional states vary regularly, and (3) how much punishment a person should receive for a given action does not vary over the times during which one might consider the case (for the same reason that led us to think in the example in the introduction that whether a prisoner in a particular case should be given parole does not vary).

There is good reason to think that externally induced emotions influence people's judgments in legal cases and in nonlegal contexts where people consider desert and punishment (Bandes and Blumenthal 2012). Goldberg et al. (1999) found that subjects primed to feel angry were more likely than other subjects to judge that individuals who committed harms should be punished harshly. They also found that angry subjects ignored intent when judging how much

punishment perpetrators should receive—only subjects' anger level predicted their judgments of the level of punishment that the perpetrator should receive. Bright and Goodman-Delahunty (2006) and Feigenson et al. (2001) also identified effects of anger on juror judgments (See also Lerner et al. 1998; Maroney 2006).

Externally induced emotions also influence several types of judgments that are likely to influence how much punishment subjects judge that a perpetrator should receive, such as judgments about whether a person is responsible or to blame for some harm, and judgments about the moral wrongness of various actions. For instance, Keltner et al. (1993) found that subjects that were angry judged an individual more responsible than impersonal circumstances for an act of harm; sad subjects judged impersonal circumstances more responsible than the individual for the act of harm. Polman et al. (2012) determined that externally induced anger led subjects to judge the actions of others more harshly and to judge their own actions less harshly than subjects who were not angry. Eskine et al. (2011) found that subjects who were induced to feel disgusted judged various actions, such as a man eating his dead dog, a congressman taking a bribe, and a person shoplifting, as more morally wrong than non-disgusted subjects did.

It is plausible that externally induced emotions such as anger and disgust affect our judgments in such a way that if we considered the same description of a case on two different occasions, experiencing different emotions, we might make different judgments about the punishment that the perpetrator should receive. Furthermore, people should be able to agree that the truth value of the proposition that the subject considers in these cases will not vary over the time in which she considers the case; we do not usually expect the appropriate punishment in a given case to vary over the course of a day, for instance, yet the judgments one makes on the case could vary over a day as one's emotions change. Thus, we have an inconsistency in the

level of variability of our attitudes on these questions and the truth value of the propositions, in the form of attitude hypervariability.

My claim here is not that emotions are biasing or distorting such that having emotions present in one's judgment formation process poses an epistemic problem; my argument does not require assumptions about whether particular causal influences are morally irrelevant or that one's process is unreliable. This is in contrast to Greene's (2008) debunking argument, which I discuss below, which involves as a *premise* that emotional processes are less reliable or that moral beliefs produced by an emotional process are less likely to be true than those produced by a nonemotional processes. [36] Rather, my argument requires only the premises that varying emotions can produce a change in judgment, that when making moral judgments about cases in ordinary life (and even in philosophy) our externally induced emotions are likely to vary, and that the truth value of the proposition in question is not varying over the time period in which we make judgments about the case. Consequently, we face an inconsistency in levels of variability in the form of attitude hypervariability, and we are thus in a poor epistemic situation with regard to propositions about appropriate punishment in particular cases.

Here, as in the case of moral beliefs that change with age, there is a natural response. Even if variation in externally induced emotion produces variation in our judgment, and our

---

[36] In his (2014) he develops a subtler argument for the unreliability of automatic processes as a result of their evolutionary history. In part his strategy consists of using less contentious normative assumptions about the moral relevance or irrelevance of certain factors to support premises about the moral relevance or irrelevance of other factors whose relevance is more contentious.

general process for making judgments about punishment is thus unreliable, one may argue that we can identify our erroneous punishment judgments by isolating the faulty part of our process. One could argue either that that one's judgments produced by a process containing externally induced emotion or anger in particular are more likely to match, or, conversely, that one's judgments formed in the absence of such emotion are more likely to match (see discussion in Feigenson 2010; Bandes and Blumenthal 2012).

Again, showing the epistemic superiority of one or the other of these is hard to do without invoking assumptions about the moral relevance of features of the cases or the reliability of moral belief processes involving emotion. On the one hand, since the emotions in question here are externally induced, they are not inspired by the case, and so they cannot be tracking some morally relevant feature of the case (they are not tracking anything about the case)—this might lead one to think that processes involving emotions are unreliable. However, it is an open possibility that we are better at making moral judgments when we are emotional than when we are unemotional. Emotions affect our processing of the case and the saliency of different features of the case. Sinnott-Armstrong (2006) has argued that emotions can adversely narrow one's attention; Allman and Woodward (2008) argue that emotions can sometimes usefully broaden one's attention. Certain emotions might put us in a heightened state and help us be more attuned to morally relevant features of the case, or may make us process information more deliberatively, or make judgments faster, leading us to disregard irrelevant details. To settle the question of whether these effects of emotion help or hinder our judgments, though, we would need to make normative claims about which features of the cases are morally relevant and which details are irrelevant. Anger in particular appears to lead to more extensive processing in cases of injustice (Litvak et al. 2009). In the case of judgments about appropriate punishments, it is not implausible

to think that we make more just decisions when we are angry. In the absence of normative assumptions, the evidence here is not conclusive either for or against the claim that externally induced emotion or anger in particular help us make better moral judgments in general or in the specific case of judgments about punishment.

Thus, because of the probability that variation in emotion will produce variation in our judgments about appropriate punishment in individual cases where the appropriate level of punishment does not vary, and it is not clear whether the presence of emotion makes us more likely to make correct judgments, we should reduce confidence in our judgments about punishment.

#### 3.3.5.4 Attitude Hypovariability: The Incorrigibility of Disgust

Now I will propose a case of attitude hypovariability, where attitudes are insufficiently variable for them to be tracking moral propositions. In the simplest attitude hypovariability debunking argument, one has reason to think that the truth value of the moral proposition varies over the time period during which it would be useful to have matching attitudes about that proposition, and yet one is likely to either maintain a belief that p or lack a belief that p over the relevant period. Some possible cases might be extremely demanding or undemanding attitudes about duties that we have good reason to think are imperfect, such as maintenance of a belief in the proposition that one ought to give away large sums of money over a time period that includes time when one is not capable of giving away large sums of money; the proposition that a group of people warrants assistance over some long duration (a span of time during which one expects that people are sometimes in need and sometimes not), or that a group of people warrants some set amount of aid over a duration in which the number of people in the group varies substantially

(along the lines of Bloom's argument, which assumed that the number of people suffering makes a difference to how much assistance should be given them); or that someone is a good person (if one allows for the possibility that the person's character could change). However, I am not currently confident about any of these assumptions. So I will examine a different sort of case.

The case I am using here is the case of moral beliefs that were formed as a result of disgust (whether the disgust extraneous or inspired by features of the case or vignette itself). This case is a somewhat more complicated example of attitude hypovariability than the simplest case that I just described, because in this case the hypovariability of attitudes is not in tension with what we know about variation in the truth value of the moral proposition but rather what we know about variation in the factors that produce the disgust: though the truth value of the proposition may not vary over the relevant time period, when the factor that led one to feel disgust toward the object varies, one's disgust reaction is unlikely to vary, meaning that, *once activated*, one's disgust is not sensitive with regard to that factor—disgust is still likely to be active in the absence of the factor. If disgust is to produce truth tracking moral judgments, whatever triggers it must be morally relevant. Even if disgust is not activated in the absence of these morally relevant factors (and so, prior to activation, may be sensitive), and even if it is activated in the presence of these morally relevant factors (and so may be adherent), once activated, disgust is no longer sensitive to that factor. Moral beliefs that are the product of disgust, then, cannot be trusted to be tracking the varying considerations they would presumably need to be tracking if they were tracking moral propositions.[37]

---

[37] One might not take this by itself to be reason for concern: even if one is clearly insensitive with regard to some factor, and so not tracking the truth value of the proposition, one could

As I mentioned in the previous section, there is evidence that a person's level of disgust can affect his or her moral judgments on a variety of questions (Haidt and Hersh 2001; Wheatley and Haidt 2005; Schnall et al. 2008; Horberg et al. 2009; Inbar et al. 2009; Zhong et al. 2010; Helzer and Pizarro 2011; Chapman and Anderson 2013; Jones and Fitness 2008; however, see also May 2014 and Landy and Goodwin forthcoming for objections). Two key features of disgust are the following: First, entities that we consider disgusting can contaminate other entities that become associated with them—disgustingness can be transmitted quite easily to things we did not previously consider disgusting. Second, once one associates disgust with an entity it is hard to stop thinking of that entity as disgusting even when the situation changes or when one acquires new evidence. Together, we can call these two features of disgustingness *stickiness*—disgustingness is easily transmitted between entities, and once disgustingness is attached to an entity it is hard to revise that association. To support my attitude hypovariability argument in the case of disgust, I appeal to the incorrigibility of moral beliefs influenced by disgust. In the next section, I will construct a different debunking argument that appeals to the transmissibility feature of disgust.

Classic cases that support the stickiness of disgust judgments include subjects' unwillingness to drink juice stirred with a new comb or a sterilized cockroach, and unwillingness

---

nonetheless be quite confident that the proposition is true (as one believes) and that the truth value of the proposition varies little, and so one could simply be grateful that one happens to have a matching attitude about that proposition. However, disgust also has a property of transmissibility that I will use in the next section to develop a distinct debunking argument that should undermine one's confidence that the proposition is true.

to wear the clothes of an amputee (Rozin et al. 1994). One might want to object that these are nonmoral contexts, and that perhaps disgust is not sticky in the same way in moral contexts. However, there is reason to think that the stickiness of disgust extends to putatively moral contexts as well. On the incorrigibility side, Haidt's work on moral dumbfounding suggests the difficulty people have in revising both their disgust response and their resulting moral judgments about a variety of cases, such as ostensibly harm-free incest (Haidt, Bjorklund, and Murphy 1999). On the transmissibility side, studies showed hesitancy on the part of subjects to wear clothing that previously belonged to a convicted murderer (Rozin et al. 1994).

Admittedly, there is some evidence showing that people can overcome or moderate the influence of disgust on their moral judgments (Haidt et al. 1993). This does not undermine the idea that people should distrust their moral beliefs influenced by disgust; it just means that some people are able to overcome the influence of disgust, which means that their moral judgments end up not being influenced by disgust. More pertinent is the evidence showing that in some contexts some people *can* revise their disgust associations, making disgust less incorrigible than I have suggested. If there is variation in disgust's stickiness across individuals, then my point is that individuals whose disgust tends to behave in a stickier manner should be more wary than those whose disgust is less sticky. Also, whether disgust exhibits stickiness in the same way across contexts may be disputed. Strohminger (2014) argues that disgust constitutes a "psychological nebula," with overlapping but diverse functions in different contexts; for that reason there may be different degrees of stickiness in different contexts. What will matter for my arguments is whether disgust is sticky within moral contexts.

I argue that the incorrigibility of disgust means that moral beliefs that are formed by a process involving a disgust reaction are too resistant to modification to be sensitive to changes in

the factors that inspire disgust to begin with. Setting aside whether any of the things that inspire disgust are actually morally wrong, for disgust to track moral wrongness it would need to be corrigible in the face of the realization that the factor that produced the disgust about the object was not in fact present or is no longer present—e.g. learning that a sweater was not worn by Hitler, or that someone did not do the evil deed they were accused of, etc. Given the difficulty we have dissociating feelings of disgust from entities and actions once we have begun to feel disgust toward them, our moral beliefs influenced by disgust cannot be tracking morally relevant features of the world. Thus, one of the properties of disgust makes it unlikely that the moral beliefs it produces are truth tracking.

This argument does not involve claims about what things should be treated as disgusting, what features of the world are morally relevant, or what sorts of things are likely to be right or wrong, which is what makes the argument unusual. It does involve denying certain types of constitutive relationship between feeling disgust toward an action and that action being wrong, but the specific constitutive relation that would be required is something few want to endorse.[38]

It is possible that there are other causes of moral beliefs that also have this property of being incorrigible that is passed on to moral beliefs and that is inconsistent with the possibility of those beliefs tracking truth. For instance, anger and contempt are often difficult to eradicate even in the face of many different types of new information—perhaps they too are inappropriately incorrigible, producing attitude hypovariability, and a similar debunking argument could be

[38] Although Kass claims that disgust can serve as a signal of moral wrongness, he does not claim that a feeling of disgust in response to something makes that thing bad or wrong. However, Prinz's view may involve this sort of constitutive relation between emotion and moral fact.

constructed for moral beliefs influenced by such emotions. Correspondingly, there may be processes that produce moral beliefs that are subsequently too easily lost to be tracking whatever activated them.

Kelly (2010, 2014) elucidated the two features of disgustingness that I am calling stickiness and used them for moral epistemological purposes, reaching the same conclusion that I am advancing, that disgust is an unreliable influence on moral beliefs. However, his argumentative strategy is different from the one I am proposing. In my view, there are four distinct arguments Kelly employs to support his conclusion that disgust is problematic in such a way that "we should disregard, discount or discredit those intuitions and be suspicious of the judgments that they influence, to the extent that we can" (Kelly 2014, 137).

One argument relies on the assumption that poisons and parasites are not morally relevant. For instance, Kelly says, "a large part of what disgust properly responds to has nothing to do with morality but is a reaction to cues likely to mark poisons and parasites" (2011, 147). Elsewhere, Kelly and Morar write, "a large number of cues that have nothing to do with highfaluting moral issues also activate disgust, namely those associated with poisons and parasites" (2014, 22). Behind these claims is an assumption about the sorts of considerations that are moral—at the least, there is an assumption that poisons and parasites are not morally relevant things to track. (One way to justify this would be if Kelly were working with the assumption that the moral facts are quite similar to the social norms that we have, which could then license this assumption that poisons and parasites are morally irrelevant.) This argument appears to be a variety of argument from morally irrelevant factors. Schafer discusses this type of argument, noting that one may conclude that a particular influence is unreliable because it responds to features of the world that we think are morally irrelevant. For instance, we may consider beliefs

cause by disgust to be an unreliable cause because it is influenced by things (cleanliness and purity) that we take to be morally irrelevant (477).[39] This strategy for assessing reliability, though, requires that we employ some normative assumptions about features of the world being morally irrelevant (even if in this case they are not the most controversial features).

A second argument suggested by Kelly's work might be interpreted as somewhat similar to the malfunctioning component argument that I introduced above, except that its method for showing that the component of the moral judgment process is unreliable at detecting what it would have to detect if it were to help one track the moral truth is to show that the component was designed for a purpose different from the one it would need to serve to promote truth tracking. Kelly provides an evolutionary story about disgust, according to which disgust evolved to help humans avoid poisons and parasites via initially distinct mechanisms that become functionally integrated (2011). Subsequently, disgust was co-opted to serve a second function, that is, to help regulate social interactions, especially by identifying group boundaries and social norms (Kelly and Morar 2014, 4). The proposal is that the disgust reaction retains many of the features that promote poison and parasite avoidance. If we assume that social norms roughly reflect the moral truths, we could conclude that moral beliefs produced by our disgust system are unreliable if the disgust system is not reliable at producing moral beliefs that match social norms. This latter claim would be supported mainly not by looking at the behavior of our disgust system and showing that the beliefs it produces regularly fail to match social norms; rather the main source of support for the claim is that because the disgust system was initially evolved for one

[39] This is similar to Prinz taking empathy to be a poor guide to moral judgments because it is not responsive to considerations that are morally relevant.

purpose it is unlikely to be as reliable at the second, distinct purpose it was put to. Here the epistemic problem is posed by the fact that our disgust faculty is tracking social norms imperfectly. On the assumption that social norms are roughly correct, this gives us a debunking argument. However, it would be better to support the claim that our disgust system does not reliably track social norms with evidence about actual mismatches between social norms and disgust-influenced moral judgments than with evidence showing that moral disgust faculties are a byproduct of a system that initially was tracking one thing and is now tracking another. This evolutionary evidence surely increases our estimation of the probability that disgust fails to track social norms but a look at how our disgust system actually behaves—a look at the proximal process we ended up with—would be a better source of evidence.

Third, Kelly argues that the disgust process is trigger sensitive or is biased toward being better safe than sorry with respect to detecting dangers, producing a high proportion of false positives; he infers from that that the system is likely to have the same bias with respect to moral judgments (2011, 147). We cannot make this inference, though, without a sense of the relation between poisons and parasites and moral properties, or a sense of the distribution of moral properties in the world.[40] (For instance, if we are surrounded by evil things that should be avoided, a hyperactive disgust system might be responding to the numerous bad things surrounding us and not producing many false positives; or if poisons and parasites are morally neutral, there is no reason to think disgust reactions is producing false positives; or if avoiding

[40] Schafer highlights the value of background knowledge about the distribution the normative properties in the world when trying to indirectly assess reliability of normative process (2010, 478).

people with pathologies is far worse than putting oneself in circumstances where one is harmed, then disgust would certainly not be a better safe than sorry guide to the sorts of actions one ought to take.) The fact that disgust is better safe than sorry with respect to dangers does not mean that it also has a better safe than sorry bias with respect to other things. A system could well be disposed to produce a high proportion of false positives with respect to one thing but produce a low proportion of false negatives with respect to a second thing. Again, if the moral facts turn out to correspond to our social norms, we could perhaps assess empirically whether disgust produces many false positives with regard to social norms or we could compare features of social norms with features of poisons and parasites to assess whether production of false positives with regard to the latter would be associated with false positives with regard to the former if the same system was used for both.

Each of these three arguments suggested in Kelly's work involves reliance on normative assumptions or is unconvincing. The attitude hypovariability argument from the incorrigibility of disgust is meant to lead us to the conclusion that we should decrease confidence in moral beliefs influenced by disgust while making fewer normative assumptions and avoid the other problems encountered above.

In this section, I have formulated a debunking argument from inconsistent variability. The argument emerges from a strategy for assessing the epistemic status of beliefs in which we compare the levels of variability between attitudes and the truth value of a proposition. Such a comparison can result in three findings—that the level of variability in one's attitudes is consistent with the level of variability we expect in the truth value of the proposition; that the level of variability in one's attitudes is higher than that in the truth value of the proposition (attitude hypervariability); or that the level of variability in one's beliefs is lower than that of the

truth value of the proposition (attitude hypovariability). The first is consistent with us being in a good epistemic position; the last two indicate that we are in a poor epistemic position. I have shown that this strategy has particular value in the moral domain, where the argument from inconsistent variability allows us to draw epistemic conclusions without making the substantial assumptions that other debunking arguments make about the moral irrelevance of causal factors or the unreliability of one's process. Applying the attitude hypervariability argument to two types of moral beliefs, I have shown that certain political beliefs are not tracking the truth because they are likely to change as one ages, and that judgments about punishment do not track the truth because they are vulnerable to influence from emotions. In each case our attitudes vary more than the truth value of the propositions, and so we do not track the truth with regard to these propositions. Applying the attitude hypovariability argument to the case of moral beliefs influenced by disgust, I argued that the incorrigibility feature of disgust makes moral beliefs that are the product of disgust too invariant to be tracking the truth of moral propositions. Consequently, we should reduce confidence in these types of beliefs.

### 3.3.6 The Inappropriately Improbable Belief Argument

The argument I discussed in the previous section, the inconsistent variability argument, has to do with whether the rates of variation of attitude and truth value of a proposition over some period of time are compatible. The epistemic problem is produced by incompatibilities in actual, observed rates of variation or anticipated levels of variation over a period of time. Information about causes can inform that type of argument by informing our estimations of probable variation in attitudes. By contrast, my third debunking argument, the inappropriately improbable belief argument, which I will now discuss, has to do with the idea that subject S, who believes p,

could have easily believed ~p, and the circumstances in which this is epistemically problematic. This argument is importantly distinct from the inconsistent variability argument, though it has some similarities. The problem arises in conditions where the truth value of the proposition of interest is static over the relevant period of time. Again, though, we do not need to make an assumption about what that truth value of the proposition is.

The inappropriately improbable belief argument offers an explication of a familiar epistemic concern: the idea that one could have easily lacked one's belief or that one's belief could have easily been different.[41] A key premise in this argument, though it is frequently not made explicit in other efforts to characterize the problem, is that in the range of circumstances in which one was likely to have a different attitude about p, the truth value of p was unlikely to have been different.

We assess the probability that we could have believed otherwise in various ways. Peer disagreement can be evidence that one could have more or less easily believed something different. Information about the causes of beliefs can also be used to inform an estimate of the probability that we will believe the proposition.

I will begin by examining the epistemic problems we can identify in a case in which one is equally likely to believe p and ~p. Then I will examine the case in which one believes p but one was unlikely to believe p, which is a situation that has attracted much attention in the debunking literature.

---

[41] This argument is closely related to discussions of what varieties of luck prevent a belief from being justified or constituting knowledge, but I will not be able to explore this relationship here. Dunaway (ms) discusses this sort of argument.

### 3.3.7 Equiprobable beliefs

Suppose S is presented with proposition p. In the range of circumstances where S is likely to form beliefs about p, p has the same truth value. Given what she knows about the causes that influence her attitude on this kind of proposition, S finds that she had a .5 probability of believing p and a .5 probability of believing ~p.[42] (Suppose that once she forms her belief about p, she retains it over the duration of interest). S can infer that she does not have a high probability of believing p given p, or a high probability of not believing p given ~p. In other words, regardless of the truth value of p, and regardless of which belief she happens to have, S can infer that she is not adherent or sensitive with regard to this proposition.

A toy case to illustrate this is that you could tell that a person's process for taking a multiple choice exam is not truth tracking, even though you do not know what the correct answer is, if you know they have an equal chance of choosing each option, when only one can be correct.

### 3.3.8 Highly improbable belief

Now consider a case where S believes p but was very unlikely to believe p. S takes herself to have an entitlement to her belief that p and she believes that she is sensitive and adherence with regard to p. Again, suppose that in the range of circumstances where S is likely to form beliefs

[42] S knows this because e.g. she is sure to use some method, and many people use this method and 50% believe p, or S has used this method for similar propositions in the past and 50% of the time she believed the proposition, etc.

about p, p has the same truth value, and that once S acquires a belief about p, she will maintain it over the time period of interest. Here I have in mind basic moral propositions, such as that pain is generally bad, or that purity is to be pursued, that are not inferred from other beliefs. Now, S discovers that although she believes p, the causal process that produced her attitude about p was highly unlikely to make her believe p. In this case, S cannot maintain that her belief is true without acknowledging that she is not adherent with regard to p.

To illustrate this argument I will return to the case of moral beliefs influenced by disgust. Earlier I highlighted that the incorrigibility feature of disgust means that moral beliefs influenced by disgust may be insensitive. Here I want to appeal to the transmissibility feature of disgustingness.

Humans are disposed to feel disgust in response to outgroups. If a person feels disgust in response to a particular outgroup, she might come to associate disgust with various traditions that the outgroup engages in—whatever those traditions are—and symbols associated with the group—whatever those symbols are—leading her to classify those traditions and symbols as wrong or to be avoided. Or, if one feels disgust at the thought of eating a particular food (e.g. squid, dogs, insects), one may come to feel disgust toward whatever group of people eats that food. As Kelly observes, "Certain activities and the people who engage in them can become disgusting, as can entire groups of people, the values and norms that bind them together, and the symbols that indicate membership in the group itself. For instance, in the United States, devoted Democrats might find distinctively Republican policies, values, and iconography disgusting, and vice versa)" (Kelly 2011, 144). Kelly and Morar discuss a case described in Henrich et al. (2010), an anthropological example about coming of age rituals in neighboring tribes, that might illustrate this phenomenon: "the Etoro believe that for a boy to achieve manhood he must ingest

the semen of his elders. This is accomplished through ritualized rites of passage that require young male initiates to fellate a senior member (Herdt 1984/1993; Kelley 1980). In contrast, the nearby Kaluli maintain that male initiation is only properly done by ritually delivering the semen through the initiate's anus, not his mouth. The Etoro revile these Kaluli practices, finding them disgusting" (61). Presumably, each group condemns the other's practices because they are performed by the outside group.

I claim that the transmissibility feature of disgustingness produces some moral beliefs that are inappropriately improbable. If S only believes that an action is morally wrong because she feels disgust toward it, and she acquired that disgust as a result of an improbable association between that action and something else toward which S felt disgust, her moral belief is improbable. For instance, if S feels disgust toward the burial practices of another group and comes to feel disgust toward and condemn as wrong a completely unrelated practice of the outgroup, such as their polygamy. Supposing that the truth value of proposition that the action is wrong is the same in the range of situations where it would be useful for S to have a matching attitude about whether the action is wrong, S's belief that p, if true, is not adherent. Some might maintain that S still has knowledge, even though she is not adherent because adherence is not required for knowledge. Even if this is so, I claim that S is in an epistemic situation that is poor in the sense that should concern us: S has some goal she wants to achieve, such as accomplishing an obligation where it would be helpful to having a matching attitude about whether it is bad to take some action. S cannot simultaneously retain her defeasible assumption that her belief is true, and acknowledge that if her belief is true it is the product of a nonadherent process.

Note that we are still talking about proximal psychological and developmental causes here: the idea is that S's psychological processes did not make her likely to believe p. Now, this

argument is different from the problem that one's belief could have easily been *false*,[43] and from the problem posed if we lack a reason to think our higher order process (evolution) was tracking anything morally relevant (what Schafer terms arguments from insensitivity) or if we have a reason to think our higher order process was tracking something morally irrelevant (e.g. fittedness to one's environment).

This argument does resemble what Schafer calls the arguments from frequency, depending on how this type of argument is understood. One way to understand the argument says that there are many different normative processes that evolution could have plausibly produced, and few possible reliable normative processes (e.g. because, without making an assumption about which moral system is correct, we assume there is just one or a few correct moral system), and so it is very unlikely that evolution would have produced a reliable one. Thus we have no more reason to think the process we ended up with is reliable than any other process we might have evolved; none of them is likely to be reliable. That argument is distinct from my inappropriately improbable belief argument, because it does not start with or address the prima facie assumption that our moral system is roughly correct. This argument certainly may be vulnerable to Schafer's second objection that it does not overcome the defeasible assumption that our normative faculty is reliable, described below.

Another interpretation of the argument is that there are many different normative processes that evolution could have produced, and it was unlikely we would end up with a

---

[43] Bedke (2014) discusses the problem posed by the range of moral facts that are conceptually possible, and he notes that Street (2008) mentions this as part of her presentation of the Darwinian Dilemma.

reliable one because we take our current normative process to be reliable, and obtaining this particular process that we have was improbable. This argument is similar to what I have called the argument from inappropriately improbable belief, but applied at a higher order, about the improbability of obtaining the proximal faculty we have. I introduced the argument as applying to a case where proximal causes make S's belief, which S initially took to be true and the product of a truth tracking process, improbable. In the inappropriately improbable belief argument we reached the conclusion that there is substantial tension in maintaining that one's belief is true and acknowledging that one is not adherent with respect to that proposition. This second interpretation of Schafer's portrayal of this skeptical argument says the evolutionary process was unlikely to produce the normative faculties that we actually have. I will advance this sort of higher order version of the argument from inappropriately improbable belief in the next chapter, when I argue that the improbability of evolving the specific fairness normative faculty that we have is in tension with the prima facie assumption that the faculty is reliable in a way that undermines our reason for thinking the faculty tracks truth.

Schafer characterizes two objections to the Arguments from Frequency—first, that evolutionary findings suggests humans were very likely to have evolved normative faculties sensitive to the things our current normative faculties respond to, such as pain and reciprocity. This is certainly an important objection to the second interpretation of the argument, and I will address it in the next chapter. The second objection is the reminder that to avoid begging the question against the realist, the skeptic has to start out assuming that we have "some sort of a priori entitlement to trust both our normative and our perceptual faculties" (482). Schafer writes, "Now, consider the set of all the possible normative faculties of any sort. Surely, in this broad class of normative faculties, most possible normative faculties will be not terribly reliable, by our

current lights. Thus, our a priori entitlement to trust our own normative faculties must amount to an entitlement to treat them as more reliable than those of an arbitrarily chosen creature. But if we have a general entitlement of this sort, why shouldn't we also be entitled to regard our normative faculties as more reliable than those of an arbitrarily chosen product of evolution? After all, this latter entitlement is simply a restricted version of the former—an entitlement that we are taking for granted here" (482-483). So, if the a priori entitlement was enough to overcome worries produced by the point that numerous other normative faculties were *possible*, the idea is that it should be enough to overcome the worry produced by the point that humans could have easily *evolved* different normative process. The difference between these two worries, though, is similar to the difference between doubting one's belief because one realizes it might have been produced by a demon and so one could have easily believed something else, and doubting one's belief because one has good empirical evidence that one's belief is (proximally) the product of a psychological process that was likely to produce a different belief, or (distally) the product of an evolutionary process that empirical evidence says was likely to produce a different moral psychology.

In this section, I have introduced characterized two ways in which the improbability of a belief may indicate that one is not tracking the truth of the proposition—a situation in one was as likely to hold the belief as not and a situation in which one was unlikely to hold the belief. For either situation to reveal an epistemic problem, one must have reason to think that truth value of the proposition is static over the relevant contexts and one will retain one's belief over the period of interest. For the second situation, it must also be the case that one's only reason for belief comes from a prima facie entitlement to hold beliefs produced by the faculty it came from or an assumption that the belief is the produce of a truth tracking process. Discovering that if one's

belief is true, then one's faculty is nonadherent undermines one's entitlement to trust beliefs produced by that faculty and thereby one's reason for believing p.[44]

## 3.4 OBJECTIONS

A general objection to these arguments is that the example psychological processes I have chosen are not the right processes to look at for an assessment of whether the moral beliefs they influence are tracking truth. Rather, it is some other process, such as the process of wide reflective equilibrium or conscious reflection, for instance, that ultimately determines whether we track moral truth. Perhaps, my opponent might suggest, processes related to disgust and sympathy merely feed into a broader process that weighs the various considerations and issues a conclusion: the disgust process might well be unreliable, but if reflective equilibrium or some other higher level process counteracts this unreliability, disgust may influence moral beliefs without being a cause for concern. Even if the empirical evidence shows that moral actions and expressed judgments are sometimes influenced by disgust, my opponent might argue in such cases the more reliable process of reflective equilibrium just failed to engage. In other words,

[44] However, my opponent might maintain that even if one's process were shown not to be truth tracking, one may nonetheless have reason to retain one's belief on the subject. If this is one's position, I think it is worth pointing out that although one will retain that true belief through the relevant time period, if this un-truth tracking process is operating in other contexts that one cares about, such as the moral beliefs of the next generation, one must acknowledge that those people are very unlikely to have true beliefs on the topic.

perhaps I have mischaracterized the relevant proximal process.

My view on this is that it is worth investigating what information we have that could bear on whether the process of reflective equilibrium (or any other candidate broader proximal process) is in fact reliable. It is also important to investigate how frequently this more reliable broader process is actually employed. If we possess this reliable process of reflective equilibrium but we employ it very rarely, then the broader process by which we reach our moral beliefs is not very reliable. We might maintain the hope that perhaps we can revise our current practice and use reflective equilibrium more frequently, but the fact would remain that as things stand we are not tracking the truth. My aim has been not to show that moral knowledge, justification, or truth tracking is not possible, but rather that we should reduce confidence in significant subsets of the moral beliefs that we currently have—namely, those types of moral beliefs that suffer from the problems that I have characterized in this chapter.

## 3.5 THE SCOPE OF DEBUNKING ARGUMENTS

In each of these arguments, information about a proximal process can be used to support the conclusion that moral beliefs obtained via that process are not truth tracking. We need not look at the evolutionary origins of disgust or sympathy to inform our conclusion that moral beliefs produced by these processes track the truth.

One reason people might be inclined to focus on information about the evolutionary origins of moral beliefs is the thought that if we could assess the reliability of evolution at producing reliable normative faculties, we could draw general conclusions about whether we track moral facts, because the influence of evolution is so wide-reaching—all moral beliefs are in

some sense caused distally by evolution. By contrast, it is reasonable to think that certain proximal causes, such as disgust and sympathy, influence only subsets of our moral beliefs. For this reason, by targeting only moral beliefs subject to influence by disgust or sympathy, the debunking arguments I have offered are of limited scope. Kelly (2014) has characterized these debunking arguments of narrow scope as "selective debunking arguments."

In principle, though, one could supply proximal debunking arguments that apply to all moral beliefs: if there are certain components that all moral beliefs production processes share, then determining that those components were unreliable would permit a global rather than selective proximal debunking argument. For instance, some have claimed that disgust is the basic moral emotion, that it does not influence purity and sanctity violations alone but rather underlies all our moral beliefs. If this were true and if my argument about the unreliability of disgust were true, we would have reason to distrust many or all of our moral beliefs. I suspect that as a matter of fact there is substantial diversity in the processes that produce different types of moral beliefs and so we will need to assess each process and thus each subset of moral beliefs separately, but this is an empirical question. At the same time, it is important to note that if we have distinct evolutionary stories for different types of moral beliefs (e.g. one evolutionary explanation for justice and fairness norms and another for purity norms), we can have distal debunking arguments that are themselves of only limited scope.[45] In the next chapter, I will formulate a distal debunking argument that applies only to beliefs about the fair distribution of resources.

[45] See Kelly (2014) for discussion of selective debunking arguments and an example of what I would characterize as an evolutionary debunking argument of limited scope.

Thus, proximal debunking arguments need not always be of narrower scope than evolutionary debunking arguments.

## 3.6 CONCLUSION

I have proposed three distinct debunking arguments that reach epistemological conclusions using information about the psychological and biological causes of moral beliefs. I illustrated each argument with a distinct type of moral belief. The properties of those proximal causes alone license conclusions about the epistemic status of the beliefs they influence: a moral belief does not track the truth if it hinges on sympathy activation, is influenced by disgust, is a certain sort of moral political belief likely to change over one's lifetime, or is a judgment about punishment that is vulnerable to the influence of emotions that are likely to vary.

Here is a summary of the relationship between the three arguments I have introduced: if we know that the truth value of a proposition is static across the conditions in which it is valuable for us to match it, then to have a high probability of matching the truth value of the proposition, one should hope to have (1) a high probability that one's attitudes about that proposition over the relevant period of time are also static, (2) a high probability that one believes p or a high probability that one does not believe p—not an equal probability, and (3) the attitude that is the probable one. If one violates (1) and one is likely to have varying attitudes about p, one can construct an attitude hypervariability argument. If one meets (1) but violates (2), so one's attitudes are likely to be static over the relevant period, but one is equally likely to believe p as ~p, one can construct an inappropriately belief argument due to equiprobability. If one meets (2), but does not meet (3), so e.g. one believes p, but one was very unlikely to believe p, one can

construct an inappropriately improbable belief argument due to inconsistency. If one knows that the truth value of a proposition is likely to vary over the time period of interest, and one knows that one's attitudes about the proposition are likely to be static over significant portions of that time period, one can construct an argument from attitude hypovariability. If one can identify the factor in the world that a component of a process would need to be picking out in order for it to contribute to the process tracking the truth of some moral proposition, and one can show that the component of the process does not in fact perform that function or does not do it sufficiently well, one can advance a malfunctioning component debunking argument.

The malfunctioning component argument employs information about a component of the causal process that produces attitudes. The inconsistent variability argument employs information about causes to estimate patterns of variation in one's attitudes about a proposition. The inappropriately improbable belief argument similarly employs information about causes to judge the probability that one would obtain the attitude one has. Thus, I have characterized three distinct ways that information about the causes of one's beliefs can inform our assessment of the epistemic status of those beliefs.

# 4.0 THE DISTAL CAUSES OF FAIRNESS NORMS

## 4.1 INTRODUCTION

In the previous chapter, I introduced the inappropriately improbable belief debunking argument. This argument says that S believes p but S was unlikely to believe p given the psychological process she uses to make judgments about propositions like p. When various other conditions hold, the improbability of this belief should lead S to reduce confidence in it. As I suggested at the end of the previous chapter, we can advance this argument at a higher order—at the developmental level, we can say that S believes p via a psychological process that makes her likely to believe p, but she was unlikely to obtain that process, developmentally. And one can construct this argument at a third level, as I will do in this chapter: even if S believes that p, was likely to believe that p given her psychological faculties, and was likely to develop those psychological faculties, one can construct a debunking argument on the basis of the fact that although humans were unlikely to have evolved those developmental and psychological processes. When certain other conditions hold, I claim that humans like S who are in this situation should reduce confidence in their belief that p.

In this chapter, I apply the inappropriately improbable *process* argument to show how information about the distal origins of fairness beliefs can contribute to our assessment of whether our beliefs about fairness norms track the truth. I argue that even if our empirical

information about the proximal psychological and developmental processes that result in intuitions about fair distribution of resources is consistent with the possibility that we track the truth of these propositions, information about their distal evolutionary origins should undermine our reason for belief. I will examine what we know about the proximal and developmental sources of these intuitions, to assess if they can support a conclusion about whether we track the truth. Then I will turn to the distal origins of these processes, and argue that humans were reasonably likely to not evolve the fairness belief processes they have. It is plausible that human beings with beliefs about fair distributions of resources could have had different sets of beliefs about how resources should be distributed. If we further assume that the truth value of any fairness norms that may exist would be the same in the different conditions where humans might have evolved different fairness belief processes, this improbability of the specific fairness processes that humans have is in tension with humans having a reason to think that they track the truth.

First I discuss the nature of the fairness intuitions that are the target of my argument. Then I describe some of the empirical research on beliefs about fairness norms in adults. I discuss some of the evidence we have on the sorts of proximal factors that influence judgments about fairness, concluding that it may be consistent with some of our fairness beliefs tracking the truth. Then I describe some research on the ontogeny of fairness beliefs, which is consistent with the possibility that Westerners track the truth. Finally, I look at what we know about the evolutionary origins of the fairness norms that humans have, as well as the conditions under which various norms related to fairness could evolve. I show that on a realist view of metaethics, this empirical evidence that the fairness process we have was unlikely to evolve undermines any prima facie assumption that our process for making fairness judgments is truth tracking. I then

discuss objections to my argument.

## 4.2 THE INAPPROPRIATELY IMPROBABLE PROCESS ARGUMENT

Here is a fictional case to illustrate the inappropriately improbable process argument: Suppose there is a society where all the members have beliefs about objects in ordinary life having a property they call "halo-ness." They cannot empirically test whether this property is present, and they cannot infer the presence of halo-ness from other properties they perceive the objects to have, but they have very strong beliefs about whether it is present in any given object. They believe that this property is inherent in the object and that most adults can reliably detect it. They treat objects with halo-ness differently than those without. They assume that their ability to detect this property is reliable, since their other belief-forming processes are usually reliable. All this seems reasonable. Then they learn via divine revelation, which is another process that they trust, that when God gave them this process for detecting halo-ness, he considered all the objects in the world and rolled dice to decide which ones their halo-detection process would pick out as having halo-ness. With this revelation the people should become much less confident that there is any such thing as halo-ness. If there is such a thing as halo-ness, it is also completely unclear which objects possess it. The improbability of the population ending up with a halo process that picks out the objects it does, given their belief that halo-ness is something inherent in objects, should undermine these people's prima facie assumption that their halo process is reliable or that they track the truth with regard to halo-ness, if there is any such thing. They should reduce confidence in their beliefs about halo-ness.

The inappropriately improbable process argument says for a belief (or domain of beliefs),

even if an individual was likely given her psychological faculties to hold the belief she does, and even if she was likely to develop those psychological faculties, she lacks reason to think that she tracks the truth and should reduce confidence in the beliefs if the following conditions hold: (1) she discovers that the population she is in was unlikely to evolve that psychological process; (2) her justification or confidence in her belief is due to a defeasible assumption that it is the product of a reliable or truth tracking process; and (3) she believes that the truth value of the proposition would not vary in the circumstances where the population would have evolved a process that produced different beliefs about the proposition. I apply this argument to basic humans intuitions about what considerations are relevant for fairly distributing resources. I will support premise (1) in subsequent sections; here I will discuss premise (2) and (3).

In policy contexts and ordinary life we frequently encounter conflicts about fairness—about how resources should be distributed within a society in which people have different needs and capacities for contributing to resource production, how wealth produced by labor should be apportioned when others own the means of production, whether drastic income inequality is ok, who should have the last slice of cake, whether one sibling should receive more stickers than another, etc. There is substantial disagreement even within Western societies on some of these questions and about what how resources should be distributed in specific scenarios. Across cultures, there are also substantial differences in how people respond to opportunities to distribute resources. However, it is possible that across cultures, people tend to treat several considerations as sometimes relevant to whether a distribution is fair. The considerations I will focus on are equality, equity, and need. My claim is that even if there is cross- and within-cultural agreement about the relevance of these considerations, and even if the proximal factors that influence judgments about fairness and developmental trajectories of these fairness

intuitions give us no reason to doubt their relevance, we should still decrease confidence in their relevance for fairness because humans could have plausibly evolved entirely different fairness processes that produced intuitions that did not count equality, or equity, or need as considerations relevant to fairness. I take it that in the conditions where we might have evolved these different fairness intuitions, realism and commonsense morality would say that the truth value of whether these three considerations are relevant for resource distribution would be the same.[46] In addition, I assume that our basic intuitions about equality, equity, and need are not inferred from other moral beliefs (admittedly a substantial assumption; see Graham, Iyer, and Meindl (n.d.) for discussion of this question), and that our reason for holding them is a defeasible assumption that they are the product of a reliable process or a process that enables us to track their truth.

I should stress that the argument of this chapter hinges on the premise that humans were not likely to acquire the particular fairness process that we have. There are other types of moral beliefs produced by processes that we may have been much more likely to acquire—it may be the case, for instance, that humans were highly likely to have a faculty that produced intuitions that harm to close relations is prima facie bad. This argument, like the arguments in the previous chapter, is a selective debunking argument.

## 4.3 FAIRNESS NORMS

By fairness norms I mean any norms about how resources should be distributed amongst

[46] Metaethical positions such as social constructivism that say that the truth value of the relevant propositions would vary in those other circumstances will not be vulnerable to this argument.

individuals—for instance, that if Person A and Person B are equally in need and A receives an unearned windfall (and there are no other agents involved), A should share some of that windfall with B; or that if A did more work than B to obtain some resources, all else being equal, more resources should be given to A; or that windfall resources should always be kept to oneself. The norm may be formulated in terms of what ought to be done, or which distribution is better, or whether a person should be punished for their distribution decision, or that a distribution being of a particular type is a reason against it. All societies seem to have norms about distributing resources—traditions and expectations about how people distribute resources in various circumstances, about punishing individuals who distribute resources in certain ways, and motivation to comply with these traditions and punish those who do not. As Hertel et al. (2002) observe, "few would doubt that people are at least motivated to appear to be fair, that they feel more comfortable with fair solutions, and that charges of unfairness are serious reproaches in many kinds of conflict negotiations" (328). I will focus on norms about the distribution of windfall goods and goods that are earned via some task,[47] and norms about what resources are owed to whom, more generally. Much of the evidence on this question comes from economic games such as the Ultimatum Game, Dictator Game, Public Goods Game, and Third-Party

[47] Interestingly, one complication is that, as Henrich et al. (2004) observe, norms about how resources should be distributed may vary according to what resource is in question; different resources are often governed by different sharing rules. This may be due to the different ways in which those different goods are typically acquired—whether they are typically obtained by chance or labor, etc. (44).

Punishment Game.[48]

[48] In the simplest Ultimatum game, Player 1 is given some quantity of resources. Player 1 must propose to give some proportion of the resource to Player 2. Player 2, who knows how much of the resource Player 1 was given and what proportion of that resource Player 1 is offering, may then accept or reject the offer. If Player 2 rejects the offer, Player 1 receives nothing. This is a one shot game. The players are usually anonymous to each other. Rational choice theory predicts that Player 1 should make the lowest offer possible and Player 2 should accept any offer. People's decisions about what to do in these games can serve as a rough indication of what the subject thinks the norms are about what ought to be done in the situation.

The public goods games may come in the form of the Voluntary Contributions game, where subjects may contribute to a pot, and a Common Pool Resources game, where subjects may refrain from removing resources from a general pot. In each of these games, the money in the pot is subsequently increased by 50% or doubled by the experimenter and redistributed to each subject equally. This game works as a measure of beliefs about fairness norms by using how much people contribute to the pot or refrain from taking from the pot as an indicator of how much of a resource they think they should contribute in a situation where the payoff structure means that the group would be best off if everyone contributed some high level of their resources, but that if everyone else is contributing at that level it would be in the individual's interest to contribute nothing. So one's decision is also influenced by how one expects the others to act, which is presumably influenced in part by one's assessment of the norms that others follow.

Norms about resource distribution interact with other sorts of norms. Haidt has proposed subdomains of morality (moral foundations theory) having to do with harm and caring, loyalty and betrayal, fairness and cheating, authority and subversion, sanctity and degradation, and liberty and oppression (Haidt and Joseph 2004). Factors that may affect judgments about whether a person should receive resources include the degree to which they helped accomplish a task, their level of need (more rarely), their status or position in society, etc. So norms about harm and caring, hierarchy and ingroup or outgroup status interact with norms about resource distribution. Reciprocity also affects distribution questions, as do norms about generosity, beneficence, and wellbeing. In addition, the question of resource distribution is intricately connected with questions of property, which does not fit neatly in Haidt's moral foundations theory; I will not be able to explore this relationship here. Most broadly, fairness might be thought to have to do with how individuals are treated in relation to each other—not just how resources are distributed but also how information is distributed, whether people are treated kindly or badly, etc., and whether punishments are administered in a fair way (Mikula et al. 1990), but I focus here just on resource

---

The Dictator game is a more direct measure than the Ultimatum or public good games of the norms that might be guiding a person distributing resources. In the dictator game, again a one-shot interaction between two people, Person 1 is given some resources and then given the opportunity to give some of those resources to Person 2. It is in the monetary interest of the giver to offer nothing. When Person 1's offer deviates from 0 it may indicate compliance with a norm about how to distribute windfall resources when one has the opportunity to share them with another person. The Third Party Punishment game is the Dictator game where a third party observes and has the opportunity to punish the Dictator, usually at a cost to the third party.

distribution.

We do not have extensive information about the proximal influences on people's fairness judgments about particular situations. Aside from the fact that many people respond to things like equality, equity, and need, there is some evidence that various of these judgments may be influenced by anger, guilt, regret, gratitude, and fear (Nelissen et al. 2011, O'Mara et al. 2011, Batson et al. 2007). Whether people are more inclined to punish proposers or compensate victims of unequal resource distributions in the Dictator game is influenced by their propensity to empathy (Leliveld et al. 2012). Some research suggests that injustice judgments are processed in one way in victims of unequal distributions and in a different way by third party bystanders: "The first, more typical of victims, is characterized by strong feelings of annoyance, disappointment, depression and/or helplessness; a pronounced sense of injustice; and a larger number of action-related remarks. The second, more typical of observers, is a rather cool and detached response to unfair events in which the major part of cognitive activity consists of reflections, evaluations and assessments" (Lupfer et al., 407). There is also evidence that people's fairness judgments tend to slightly favor themselves, which Hertel et al. (2002) propose may be more common in contexts of competitiveness than in contexts of cooperation (338). Currently, the psychological story about the proximal influences on fairness judgments is highly incomplete. My argument in this chapter, though, is that even if we ultimately find that the psychological causes of fairness judgments are consistent with the possibility that we track the truth of these questions, a distal debunking argument will lead us to conclude we should reduce confidence in the relevance of factors of equality, equity, and need for the question of how resources should be distributed. In the next section, I will argue that the empirical evidence about the development of intuitions about resource distribution in Westerners is consistent with truth

tracking. Then I will turn to my evolutionary debunking argument.

## 4.4 DEVELOPMENT OF WESTERN FAIRNESS INTUITIONS

Children are disposed from very early on to look for norms, acquire and follow norms, and punish norm violators. Children "monitor others for emotive cues of proper behavior" (Richerson and Henrich 2012, 46) and they can infer the existence of a norm simply based on observing costly behavior, or the example of others, though they are selective in determining whose behavior or explicitly expressed norms to adopt (e.g. Schmidt et al. 2012). The system by which children acquire norms appears to be very flexible in some domains, such as in the purity domain and in games, in the sense children can easily acquire a wide variety of norms (at least harmless ones) in those domains, even in the absence of any apparent reason for the norm (Rottman and Kelemen 2012, Rakoczy et al. 2008).

Although there is substantial flexibility in moral norm acquisition generally, at least for fairness norms there does seem to be a developmental pattern that Western children exhibit. Western children tend to systematically acquire certain fairness norms, or beliefs that certain considerations are relevant for fairness. These children note the existence of and acquire dislike for inequality involving themselves before the age of three (LoBue et al. 2009), begin to acquire beliefs related to others' property at around three (Rossano et al. 2011; see also Friedman and Neary 2008), and begin to talk about fairness as if it is a reason that counts for or against certain proposed distributions at age 5 or 6; they defend their objection to unequal distributions by appealing to fairness considerations that they presumably expect others to respond to (discussed

in LoBue et al. 2009, 156).[49] As they develop, they sequentially acquire certain norms and beliefs in the relevance of certain considerations. Early on, children follow a simple equality norm in cases not involving themselves. When given the opportunity to distribute goods to others, for instance distributing stickers to dolls, they tend to distribute the goods equally (Schmidt et al. 2012). When dividing unearned goods between themselves and others, they appear to be guided by both self-interest and equality concerns. They keep most of the resource themselves up until ages 6 or 7, giving more to their partner as they age, though Warneken et al. (2011) found that three-year-olds shared *earned* resources mostly equally. Their level of giving in contexts like the Dictator game also tends to increase over time (Richerson and Henrich, 46).

When distributing rewards for work done, at age 5-6, children give rewards equally, ignoring quantity of work done (Larsen and Kellogg 1974; Lerner 1974). Later, they start to treat amount of work performed as relevant—they give somewhat more to those who worked more, but their distribution is not strictly proportional to amount of work done; only as adolescents does strict proportionality occur (see McCrink et al. 2010 for discussion). An age entitlement rule, in which the child gives more resources to older or bigger individuals, has also been observed among some four and five year olds, but then disappears (Thomson and Jones 2005). (One might wonder whether this is related to recognition of greater need, deference to higher status individuals warranting more, or a factor distinct from fairness, such as recognition that the older person could coerce one to give them more resources). At some point after age 5 they acquire a belief that those who contributed more to a task that produced the resource may or

[49] Children also spontaneously share toys from eight months, but it is not at all clear that this is some sort of precursor to fairness norms (LoBue et al. 2009, 155).

should receive more than those that did less (e.g. Thomson and Jones 2005 says this happens around age 7). Some eight year olds have been found to apply a benevolence rule in favor of those that are financially needy (see discussion in Thomson and Jones 2005). In older children, distribution may also be influenced by whether recipients are friends or strangers (McGillicuddy-DeLisi et al. 1994) Consideration of a person's capacity to contribute to a joint task becomes common only much later: a majority of ninth graders and college students in one study took physical disability into consideration when distributing payment to three siblings who worked together on a task—those concerned with equity seemed to take into account that the disabled child could not contribute as much to the task as the others (Thomson and Jones 2005; see also McCrink et al 2010).

This is a quick sketch of some of the evidence on the development of fairness norms in Westerners, but the existence of this kind of pattern of norm acquisition could be consistent with Western populations tracking the truth of fair distribution of resources. Individuals in the West systematically develop beliefs about the relevance of a set of considerations for the question of whether particular resource distributions are fair. Among these considerations are equality—equal distribution to individuals involved; equity—distribution proportional to amount of work put in to obtain the resource; and need. Other potentially relevant factors are self-interest, one's relationship to others, people's capacity to help obtain a resource, and some sort of age consideration (that might reflect need or hierarchy factors). Supposing that whether these factors are relevant for fairness is invariant over individuals within the culture, the fact that individuals systematically acquire concern for a set of factors relevant to fairness judgments is consistent with the possibility that individuals in the West track the truth. On the other hand, it is not knockdown evidence that these people *do* track the truth, so it is worth looking at the process by

which they obtained this developmental process to see how it bears on truth tracking.

Less data is available about the proximal influences on and development of norms about resource distribution in non-Western cultures. However, because the norms that are the target of this chapter are not specific norms about how resources should be distributed in a particular case but rather the idea that equality, equity, and need are sometimes relevant to fairness, I think it is reasonable to start with the assumption that in most cultures notions about fairness involve these factors.

## 4.5 EVOLUTION OF FAIRNESS INTUITIONS

Now I will support the claim that human populations could have plausibly evolved a capacity and inclination to make judgments about whether distributions of resources are fair that involved different intuitions about the nature of fairness. Among other things, we could have been inclined to treat need or equity as irrelevant to fairness or could have treated anything less than hypergenerous, selfless distributions as unfair.

The evidence about the probability that our fairness process could have been one that produces different intuitions comes in part from what we know about the conditions in which these kinds of norms are likely to emerge in or spread through a population and what we know about the evolutionary functions that beliefs in these norms may serve. I should emphasize that by the evolution of fairness norms, the process I have in mind is evolution understood to include socio-cultural change and multiple modes of inheritance—not just genetic. So change in human fairness developmental processes could occur over generations via changes to e.g. systems of rules transmitted by instruction and imitation, learned emotional reactions to certain resource

distributions, dispositions to dislike certain distribution schemes independent of any imitation, individual or cultural inclinations to view the world as just, capacity for empathy-development in various childhood environments, etc.

I will highlight three sources of evidence to support the claim that people could have evolved fairness intuitions that conflict with those observed in Western populations: (1) that other norms about distribution reasonably likely to come into being are sustainable in a population, (2) that norms inconsistent with the norms identified as characteristic of the West are systematically maintained among minorities in Western populations, and that frequency-dependent selection and selection for multiple norms in a population may plausibly apply to fairness norms, and (3) supposing the evolutionary function of fairness norms is to facilitate cooperation, we could have had cooperation without our fairness norms, because there are a variety of mechanisms that can promote cooperation.

I will begin with the point that other norms governing resource distribution were reasonably likely to come into being and could have been sustained in the population. In some conditions, selection may favor the unequal distribution of power and resources, making selfishness in leaders more advantageous. Makowsky and Smaldino (2014) developed a model involving multilevel selection and interaction between groups found that in conditions of substantial intergroup conflict, less democratic, more hierarchical groups with selfish leaders outcompeted groups where power was distributed more equally. In such contexts, there could be selection pressure in favor of a norm system for resource distribution driven primarily to ensure that the leader or dominant member of the group receives sufficient resources, regardless of the effects this has on levels of equality, equity, or need of non-leaders.

Boyd and Richerson (2005) and others have shown that the mechanisms that support

prosocial norms, such as reputation, punishment, and signaling, can stabilize norms that are selectively neutral or disadvantageous for the individual or the group (see also Richerson and Henrich 2012). Boyd and Richerson (2005) observe, "A problem with the multiplicity of proximal mechanisms for sustaining equilibria is that the different mechanisms and equilibria likely exhibit a wide range of functionality. We can imagine several mechanisms by which dysfunctional norms/institutions can become established. Exogenous changes may make an institution obsolete, yet it may be sustained by the mechanisms we have reviewed. … Predatory elites and other self-interested subgroups with some form of coercive power may be able to establish equilibria that disproportionately benefit them" (50). Because the systems supporting fairness norms have this flexibility, they are capable of supporting systems of norms that contradict current Western fairness norms. Once they come into existence in a population, norms other than the ones we have may be sustainable.

Even if our environmental circumstances were not substantially different, we might have evolved a different set of norms because multiple systems of resource distribution may be supportable in a single set of circumstances. In their crosscultural studies on fairness norms using Dictator, Ultimatum, and Third-Party Punishment games, Henrich et al. (2010) found that responses to games were generally correlated with level of market integration, but some economically similar societies in their study appear to have very different principles—e.g. Au and Tsimane societies have a similarly low level of market integration, but the Tsimane offered much less on average in the Dictator game than the Au.

Indeed, in many models used to investigate how certain equality norms or behaviors might have come about, there are multiple norm equilibria. As part of an effort to explain the existence of norms favoring equal division of resources, Zollman (2008) argues that complexity

of environment—specifically a hybrid of a Nash bargaining game and the ultimatum game—is one reason an equality norm may be favored.[50] Past literature offers models in which Nash bargaining games produce equal divisions of goods, but few in which the ultimatum game successfully results in equal divisions of goods. Zollman (2008) found that in some games that combined Nash bargaining games and the ultimatum game, equal divisions are more likely to evolve than in either of the component games. In the Nash bargaining game, one finds a fair 1/2, 1/2 equilibrium, a "nonfair" equilibrium, and "hyperfair" equilibria. If our fairness norms originated in these conditions, humans could have easily been disposed to divide resources heavily in favor of the other person or heavily in their own favor.

Secondly, I claim that there is sustained disagreement about fairness norms in the West—for instance, about whether equity or need matters in a given case of resource distribution—and this diversity within the population suggests the possibility of an entire human population being unconcerned with considerations like equity or need when it comes to fairness. Evidence of this deep disagreement comes from research on conservatives, especially libertarians, and liberals, different fairness judgments produced by proself and prosocial subjects, and the diversity of responses within Western populations to economic games. In studies on Western university students, the mode in a public goods game was at 0 contributions, and a secondary mode was at

[50] In a Nash bargaining game, two subjects are given a windfall to divide. Each proposes a proportion of the resource she wishes to keep. If the amount the subjects wanted to keep exceeds the total quantity of resources, neither receives anything; if the amounts the subjects wanted to keep is less than the total quantity of resources, each receives what she requested.

100% contributions (Henrich et al, 22).[51] Stouten et al. (2005) found in a group of Western subjects that the subjects more concerned with their own outcomes (proself subjects) were more likely to support equality of contributions out of a concern for efficiency than subjects more concerned with society (prosocial subjects), who were more likely to support equality of contributions because to do otherwise would be "unfair"—only the latter group continued to support equal contributions in a situation where it was not required for efficiency.

Third, humans could be cooperative without the norms about resources distribution that Westerners currently have. Many people have suggested that sociality and cooperation are key features of being human; if different resources distribution norms prevented sociality and cooperation, one might object to my argument by saying that it would not be quite right to say that humans could have easily had different resources distribution norms, because in doing so we would lose a key component of humanity. However, there are a variety of mechanisms that can

---

[51] Across cultures, there is also substantial diversity: Henrich et al. report, "The Machiguenga, for example, have a mode at full defection, but lack any fully cooperative contributions—which yields a mean contribution of 22 percent. Both the Ache and Tsimane experiments yielded means between 40 and 60 percent, like folks from industrialized societies, but, unlike industrial societies, they show unimodal distributions with peaks at 50 and 66.7 percent, respectively. Their distributions resemble *inverted* American distributions with few or no contributions at full free riding and full cooperation. Like the Ache and Tsimane, the Onna and Huinca have modes near the center of the distribution, at 40 and 50 percent respectively, but they also show secondary peaks at full cooperation (100 percent)-and no contributions at full defection" (Henrich et al. 2010, 22).

enable cooperation. According to Richerson and Henrich (2012), "In recent years, theorists have discovered many mechanisms that can stabilize cooperation (Chudek, Zhao, and Henrich forthcoming). Various forms of punishment, ostracism of non-cooperators, assortative formation of groups with like propensities to cooperate, cooperation provided as a costly signal in a mating game, and other situations in which payoffs in one game are linked to another by reputations are among the plausible mechanisms that have been studied" (66). Several of these could support the evolution of cooperation without anyone believing equality, equity, or need is relevant for the fair division of goods. Another possibility is that we could have been motivated to act in a way that is consistent with a set of fairness rules similar to our own but to do it on the basis of norms about resource distribution that conflict with the relevance of equality, equity and need. Forber and Smead (2014) have shown that it is possible for a behavior of offering an even split of resources to spread in a population as a result of spiteful behavior in an ultimatum game. It is plausible that it could be advantageous for individual humans to cooperate even without norms favoring equal or equitable distribution of goods because even the weakest members of society are better off cooperating and getting scraps than going off on their own. The many contexts in human history where cooperation actually *was* forced by violence or drastic power imbalances likewise support the idea that we could easily be cooperative and social without thinking equality, equity or need are relevant to how resources should be distributed.

Furthermore, I claim that in these alternative scenarios where humans might have had societies with different sets of norms about the fair distribution of resources, the most likely way that these norm systems would be supported is by people having correspondingly different basic intuitions about fairness. Given that much of human behavior is motivated by moral beliefs, and given the flexibility of the human norm acquisition system, (and people's willingness even now

to endorse moral rules that work against their individual interests), it is reasonable to think these alternative systems of resource distribution would be maintained and propagated by human moral beliefs about how goods should and should not be distributed. In sum, it is plausible that Westerner's fairness beliefs could be different if evolution had proceeded slightly differently—and indeed they may be expected to be different in the future when conditions change. The fairness norms that are assumed to prevail in the West—concern for equality, equity, and need—are improbable.

## 4.6 OBJECTIONS

In the previous chapter, I mentioned that Schafer (2010) raises an objection to evolutionary debunking arguments that rely on the premise that our moral intuitions were unlikely to evolve. The objection denies this premise, and says that evolution was actually likely to produce many of our moral beliefs: "the current state of evolutionary theory supports the claim that we would have developed normative faculties that are sensitive to factors like our pain and the pain of those related to us, beginning from a wide range of initial conditions. And it supports a similar claim about the range of initial conditions from which we would have developed normative faculties that were sensitive to issues of reciprocity. But so long as we are sensitive to these sorts of factors, and have our general capacities for normative reasoning and reflection, we stand a good chance of arriving at normative judgments that are reasonably reliable, at least by our present lights" (482). Similarly, White writes, "natural selection is likely to favor creatures with a sense of obligation toward their offspring" (2010, 589). I will not deny that there may be some types of moral intuitions that humans were likely to evolve in a range of conditions, but my

claim is that we were reasonably unlikely to evolve the particular intuitions about resource distribution that we have. And even if other types of moral intuitions were likely to evolve, I am inclined to think that it may not be possible to reason our way from one type of moral intuitions to others (e.g. we might have intuitions about purity without being able to reason from those to intuitions about harm; we may need basic intuitions about harm in order to infer the rest of our harm norms). Even if we had norms about harm, purity, and authority, I do not think we could infer all of our norms about fair distribution of resources.

## 4.7 CONCLUSION

The evidence about the proximal causes and developmental trajectory of fairness intuitions (at least, for Westerners) may be consistent with the possibility of tracking the truth. However, a look at the evolutionary origins of the psychological and developmental processes that produce intuitions about how resources should be distributed suggest that human populations could relatively easily have beliefs about fair resource distribution that do not take into consideration equality, equity, or need. In combination with the idea that the relevance of equality, equity, or need for fairness would not be different in those other circumstances, and that our reason for believing that these three factors matter for fairness is a defeasible assumption that the process producing these intuitions is truth tracking, the finding that humans were not particularly likely to obtain these views about fairness should lead us to reduce confidence in our belief that these factors are relevant for fairness.

# 5.0 CONCLUSION

I have argued that in the moral domain, when we investigate the epistemic status of our beliefs, we should be concerned primarily with whether we track the truth of moral propositions. If there are any obligations that humans have, there are certain propositions related to the nature of those obligations and the means of achieving them that would be valuable to match—having matching attitudes about these propositions at the appropriate times would make us more likely to complete our obligations. An important task for the future will be to spell out in further detail methods for identifying such propositions. If humans do have obligations, we should do what we can to improve our chances of matching relevant propositions at appropriate times. For this reason, it is worthwhile to investigate whether our current moral beliefs track truth.

Empirical research on the causes of moral beliefs offers us information relevant for assessing whether we track moral truths. I have characterized three types of debunking arguments that can employ information about the causes of beliefs. These are the malfunctioning component argument, the argument from inconsistent variability, the inappropriately improbable belief argument, and its higher order variant, the inappropriately improbable process argument. Each is a selective debunking argument that may be applied to subsets of moral beliefs. The malfunctioning component argument, the argument from inconsistent variability, the inappropriately improbable belief arguments can lead to the conclusion that an agent does not track the truth of a proposition. The inappropriately improbable process argument can undermine

one's reasons for thinking one tracks truth. These arguments are distinct from existing etiological debunking arguments in the literature. Furthermore, they permit us to assess the epistemic status of types of moral beliefs without appealing to assumptions about the moral irrelevance of certain causes or the unreliability of certain processes. Instead, the arguments can lead us to the conclusion that a factor is morally irrelevant or that a process is unreliable.

I developed versions of these arguments for several types of moral beliefs. I concluded that we do not track the truth of moral beliefs that hinge on sympathy, moral beliefs influenced by disgust, certain moral political beliefs, and beliefs about punishment that are vulnerable to the influence of extraneous emotions. For each of these types of moral beliefs, the conclusion about epistemic status was the product of information about proximal rather than causes. Lastly, I argued that, given certain assumptions, we lack a reason to think that we track the truth about the relevance of equality, equity, and need for the fair distribution of resources.

Each of the debunking arguments that I developed may be applied to other types of moral beliefs. Other types of moral beliefs worth investigating are beliefs related to authority, respect, hierarchy, and liberty, loyalty, deception, and property. Other types of causes to consider in the future when assessing whether moral beliefs track truth are conscious reasoning, unconscious application of moral rules, reflective equilibrium, other emotions, such as anger and envy, and the evolutionary origins of other types of moral beliefs. Because the arguments that I have discussed are selective, though, there remains the possibility that some types of moral beliefs will not be undermined by etiological debunking.

## BIBLIOGRAPHY


Allman, John, and James Woodward. 2008. “What are moral intuitions and why should we care about them? A neurobiological perspective.” *Philosophical Issues* 18:164–185.

Alston, William P. 2005. *Beyond "Justification": Dimensions Of Epistemic Evaluation*, Ithaca, N.Y.: Cornell University Press.

Armstrong, Joel B. 2013. *Tough but fair: The moderating effects of target status on the relation between social dominance orientation and fairness*. Masters Thesis. University of Western Ontario London.

Bandes, Susan A., and Jeremy A. Blumenthal. 2012. “Emotion and the Law.” *Annual Review of Law and Social Science* 8:161–181.

Bartneck, Christop and Jun Hu. 2008. “Exploring the Abuse of Robots.” *Interaction Studies* 9:415–433.

Batson, C. D., Kennedy, C. L., Nord, L.-A., Stocks, E. L., Fleming, D. A., Marzette, C. M., et al. 2007. “Anger at unfairness: is it moral outrage?” *European Journal of Social Psychology*, 37:1272–1285.

Bedke, Matthew. 2009. “Intuitive Non-Naturalism Meets Cosmic Coincidence. *Pacific Philosophical Quarterly*, 90:188–209.

Bedke, Matthew. 2014. “No coincidence.” In ed., Russ Shafer-Landau, *Oxford Studies in Metaphysics*, 9:102–105.

Berker, Selim. 2009. “The normative insignificance of neuroscience.” *Philosophy & Public Affairs* 37:293–329.

Bloom, Paul. 2013. “The Baby in the Well: The Case Against Empathy.” *The New Yorker*. May 20, 2013.

Bright, DA and J. Goodman-Delahunty. 2006. “Gruesome evidence and emotion: anger, blame, and jury decisionmaking.” *Law and Human Behavior* 30:183–202.

Brosnan, Kevin. 2011. “Do the evolutionary origins of our moral beliefs undermine moral knowledge?” *Biology and Philosophy* 26:51–64.

Chapman, Hanah A. and Adam K. Anderson. 2013. “Things Rank and Gross in Nature: A Review and Synthesis of Moral Disgust.” *Psychological Bulletin* 139:300–327.

Cikara, Mina and Susan T. Fiske. “Bounded Empathy: Neural Responses to Outgroup Targets’ (Mis)fortunes.” *Journal of Cognitive Neuroscience* 23:3791–3803.

Clark-Doane, Justin. 2012. “Morality and Mathematics: The Evolutionary Challenge.” *Ethics* 122:313–340.

Cohen, G.A. 2000. *If You're an Egalitarian, How Come You're So Rich?* Cambridge, MA: Harvard University Press.

Colaço, David, Wesley Buckwalter, Stephen Stich, and Edouard Machery. 2014. “Epistemic intuitions in fake-barn thought experiments.” *Episteme* 11:199–212.

Conway, Paul, and Bertram Gawronski. 2013. “Deontological and utilitarian inclinations in moral decision making: A process dissociation approach.” *Journal of Personality and Social Psychology* 104:216–235.

Copp, David. 2008. “Darwinian Skepticism about Moral Realism.” *Philosophical Issues* 18:186–206.

Côté, Stéphane, Paul K. Piff, and Robb Willer. 2012. "For whom do the ends justify the means? Social class and utilitarian moral judgment." *Journal of Personality and Social Psychology* 104:490–503.

Cutler, Stephen J., and Robert L. Kaufman. 1975. "Cohort changes in political attitudes: Tolerance of ideological nonconformity." *Public Opinion Quarterly* 39:69–81.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous factors in judicial decisions." *Proceedings of the National Academy of Sciences of the United States of America* 108:6889–6892.

Dunaway, Billy. ms. "Luck: Evolutionary and Epistemic." *www-personal.umich.edu/~dunaway/EvoLuck.pdf. Accessed March 23, 2015.*

Enoch, David. 2010. "The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope With It." *Philosophical Studies* 148:413–438.

Eskine, Kendall J., Natalie A. Kacinik, and Jesse J. Prinz. 2011. "A Bad Taste in the Mouth Gustatory Disgust Influences Moral Judgment." *Psychological Science* 22:295–299.

Eysenck, H. J. 1975. "The structure of social attitudes." *British Journal of Social and Clinical Psychology* 14:323–331.

Feather, N. T. 1978. "Family resemblance in conservatism: Are daughters more similar to parents than sons are?" *Journal of Personality* 46:260–278.

Feigenson Neal, Park J, Salovey P. 2001. "The role of emotions in comparative negligence judgments." *Journal of Applied Social Psychology* 31:576–603.

Feigenson, Neal. 2010. "Emotional influences on judgments of legal blame: How they happen, whether they should, and what to do about it." Nebraska Symposium on Motivation. 56:45–96. New York: Springer.

Feldman, Richard. 2006. “Epistemological Puzzles About Disagreement.” In Stephen Hetherington, ed. *Epistemology Futures.* Oxford University Press, Oxford, 216–236.

Fiala, Brian, Adam Arico, and Shaun Nichols. 2014. “You, Robot.” In *Current Controversies in Experimental Philosophy*, ed. Edouard Machery and Elizabeth O’Neill. New York: Routledge, 31–47.

Fraser, Benjamin James. 2014. “Evolutionary debunking arguments and the reliability of moral cognition.” *Philosophical Studies* 168:457–473.

Friedman, O., & Neary, K. R. 2008. Determining who owns what: Do children infer ownership from first possession? *Cognition,* 107:829–849.

Glenn, N. D. 1974. “Aging and conservatism.” *Annals of the Academy of Political and Social Science* 415:176–186.

Glisky, E. L. 2007. “Changes in Cognitive Function in Human Aging.” In Riddle D. R., ed., *Brain Aging: Models, Methods, and Mechanisms*. Boca Raton, FL: CRC Press.

Goldberg, J. H., Lerner, J. S., and Tetlock, P. E. 1999. “Rage and reason: the psychology of the intuitive prosecutor.” *European Journal of Social Psychology* 29:781–795.

Goldman, Alvin. 2014. “Naturalizing Metaphysics with the Help of Cognitive Science.” In eds., Karen Bennett and Dean W. Zimmerman, *Oxford Studies in Metaphysics*, 9:171–213.

Goldman, Alvin. 1986. *Epistemology and Cognition*. Cambridge, MA: Harvard University Press.

Graham, Jesse, Ravi Iyer, and Peter Meindl. n.d. “The Psychology of Economic Ideology: Emotion, Motivation, and Moral Intuition.” *Demos*. *www.demos.org/sites/default/files/publications/Graham.pdf*

Grant, M. J., Ross, A. S., Button, C. M., Hannah, T. E., & Hoskins, R. 2001. “Attitudes and stereotypes about attitudes across the lifespan.” *Social Behavior and Personality: An*

*International Journal* 29:749–762.

Greene, Joshua. 2008. "The Secret Joke of Kant's Soul." In *Moral Psychology: Volume 3. The Neuroscience of Morality: Emotion, Brain Disorder and Development*, ed. Walter Sinnott-Armstrong, 35–80. Cambridge, MA: MIT Press.

Greene, Joshua. 2014. "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics." *Ethics* 124:695–726.

Griffiths, Paul, and John Wilkins. n.d. "*When do evolutionary explanations of belief debunk belief?"* Online: http://philsci-archive.pitt.edu/5314/

Haidt, Jonathan and Craig Joseph. 2004. "Intuitive ethics: how innately prepared intuitions generate culturally variable virtues." *Daedalus* 133:55–66.

Haidt, Jonathan, and Matthew A. Hersh. 2001. "Sexual Morality: The Cultures and Emotions of Conservatives and Liberals." *Journal of Applied Social Psychology* 31:191–221.

Haidt, Jonathan, Fredrik Bjorklund, and Scott Murphy. 2000. "Moral Dumbfounding: When Intuition Finds No Reason." Unpublished manuscript, University of Virginia.

Harman, Gilbert. 1986. "Moral explanations of natural facts: Can moral claims be tested against moral reality?" In *Spindel Conference: Moral Realism* (*Southern Journal of Philosophy suppl. 24*), ed. N. Gillespie, 57–68.

Harris, Sam. 2014. "Forum: Against Empathy" *Boston Review*. Aug 26. http://www.bostonreview.net/forum/against-empathy/sam-harris-response-against-empathy-harris

Helzer, Erik G., and David A. Pizarro. 2011. "Dirty Liberals!: Reminders of Physical Cleanliness Influence Moral and Political Attitudes." *Psychological Science* 22:517–522.

Henningham, J. P. 1996. "A 12-item scale of social conservatism." *Personality and Individual*

*Differences* 20:517–519.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. 2004. "Foundation of Human sociality". In Henrich, Joseph, et al. *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford University Press.

Hertel, G., Aarts, H., & Zeelenberg, M. 2002. "What do you think is 'fair'? Effects of ingroup norms and outcome control on fairness judgments." *European Journal of Social Psychology*, 32:327–341.

Hewitt, J. K., J. J. Eysenck, and L. J. Eaves, L. J. 1977. "Structure of social attitudes after 25 years: A replication." *Psychological Reports* 40:183–l88.

Horberg, E. J., Christopher Oveis, Dacher Keltner, and Adam B. Cohen. 2009. "Disgust and the Moralization of Purity." *Journal of Personality and Social Psychology* 97:963–976.

Inbar, Yoel, David A. Pizarro, Joshua Knobe, and Paul Bloom. 2009. "Disgust Sensitivity Predicts Intuitive Disapproval of Gays." *Emotion* 9:435–439.

Jones, Andrew, and Julie Fitness. 2008. "Moral Hypervigilance: the Influence of Disgust Sensitivity in the Moral Domain." *Emotion* 8:613–627.

Joyce, Richard. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.

Joyce, Richard. 2006. *The Evolution of Morality*. Cambridge, MA: The MIT Press.

Kahane, Guy, and Nicholas Shackel. 2010. "Methodological Issues in the Neuroscience of Moral Judgement." *Mind and Language* 25:561–582.

Kahane, Guy. 2011. "Evolutionary debunking arguments." *Nous* 45:103–125.

Kahane, Guy. 2012. "On the Wrong Track: Process and Content in Moral Psychology." *Mind and Language* 27:519–545.

Kahane, Guy. 2013. “The armchair and the trolley: an argument for experimental ethics.” *Philosophical Studies* 162:421–445.

Kass, Leon. 1997. “The Wisdom of Repugnance.” *The New Republic*. Vol. 216 Issue 22. June 2.

Kelly, Daniel, and Nicolae Morar. 2014. “Against the Yuck Factor: On the Ideal Role of Disgust in Society.” *Utilitas* 26:153–177.

Kelly, Daniel. 2011. *Yuck! The Nature and Moral Significance of Disgust*. Cambridge, MA: MIT Press.

Kelly, Daniel. 2013. “Moral Disgust and The Tribal Instincts Hypothesis,” in *Cooperation and Its Evolution*, Eds. K. Sterelny, R. Joyce, Calcott, B, & B. Fraser. Cambridge, MA: The MIT Press, 503–524.

Kelly, Daniel. 2014. “Selective Debunking Arguments, Folk Psychology and Empirical Moral Psychology,” In *Advances in Experimental Moral Psychology*. ed. J.C. Wright and H. Sarkissian. New York: Continuum Press, 130–147.

Kelly, Thomas. 2010. “Peer Disagreement and Higher Order Evidence.” In *Social Epistemology: Essential Readings*, eds., Alvin I. Goldman and Dennis Whitcomb, 183–217. New York: Oxford University Press.

Keltner, Dacher, Phoebe C. Ellsworth, and Kari Edwards. 1993. “Beyond simple pessimism: effects of sadness and anger on social perception.” *Journal of Personality and Social Psychology* 64:740–752.

Konty, Mark A. and Charlotte Chorn Dunham. 1997. “Differences in value and attitude change over the life course.” *Sociological Spectrum* 17:177–197.

Landy, Justin F. and Geoffrey P. Goodwin. Forthcoming. “Does Incidental Disgust Amplify Moral Judgment? A Meta-Analytic Review of Experimental Evidence*” Perspectives on Psychological Science.*

Larsen, Gary Y., and Jeffrey Kellogg. 1974. “A developmental study of the relation between conservation and sharing behavior.” *Child Development* 45:849–851.

Leliveld, M. C., Dijk, E., & Beest, I. 2012. “Punishing and compensating others at your own expense: The role of empathic concern on reactions to distributive injustice.” *European Journal of Social Psychology*, 42:135–140.

Lerner, Melvin J. 1974. “The justice motive: ‘Equity’ and ‘parity’ among children.” *Journal of Personality and Social Psychology* 29:539–550.

Lerner, J. S., J. H. Goldberg, and P. E. Tetlock. 1998. “Sober second thought: The effects of accountability, anger and authoritarianism on attributions of responsibility.” *Personality and Social Psychology Bulletin* 24:563–574.

Levenson, Robert W. 2000. Expressive, physiological, and subjective changes in emotion across adulthood. In Qualls, Sara Honn and Norman Abeles, eds, *Psychology and the aging revolution: How we adapt to longer life*. Washington, DC, US: American Psychological Association, 123-140.

Litvak, P., Lerner, J. S., Tiedens, L. Z., & Shonk, K. 2010. “Fuel in the fire: How anger impacts judgment and decision making.” In *The Handbook of Anger*, ed., M. Potegal. New York: Springer, 287–310.

LoBue, V., Nishida, T., Chiong, C., Deloache, J. S., & Haidt, J. 2010. When Getting Something Good is Bad: Even Three-year-olds React to Inequality. *Social Development* 20:154–170.

Lupfer, M. B. and J. P. Rosenberg. 1983. “Differences in adults’ political orientations as a

function of age." *Psychology* 119:125–133.

Lupfer, Michael B., et al. 2000. "Folk conceptions of fairness and unfairness." *European Journal of Social Psychology* 30:405-428.

Machery, Edouard, and Ron Mallon. 2010. "Evolution of Morality." In *The Moral Psychology Handbook*, ed. John Doris. Oxford: Oxford University Press, 3–46.

Maibom, Heidi. 2012. "In a different voice." *Neurofeminism: Issues at the Intersection of Feminist Theory and Cognitive Science*. Basingstoke, Hampshire: Palgrave Macmillan, 56–72.

Makowsky, M. D., and Smaldino, P. E. 2014. "*The Evolution of Power and the Divergence of Cooperative Norms*." Available at SSRN 2407245.

Mallon, Ron and John Doris. 2013. "The Science of Ethics." In *The Blackwell Guide to Ethical Theory*. ed. Hugh LaFollette and Ingmar Persson, 169–196. Oxford: Blackwell.

Maroney, TA. 2006. "Law and emotion: a proposed taxonomy of an emerging field." *Law and Human Behavior* 30:119–142.

Marwell, Gerald, Michael T. Aiken, and N. J. Demerath. 1987. "The persistence of political attitudes among 1960s civil rights activists." *Public Opinion Quarterly* 51:359–375.

May, Joshua. 2014. "Does Disgust Influence Moral Judgment?" *Australasian Journal of Philosophy* 92:125–141.

McCrink, K., Bloom, P., and Santos, L. R. 2010. "Children's and adults' judgments of equitable resource distributions." *Developmental Science*, 13:37–45.

McGillicuddy-de Lisi, Ann V., Cynthia Watkins and Andrew J. Vinchur. 1994. "The effect of relationship on children's distributive justice reasoning." *Child Development* 65:1694–1700.

Mikula, G., Petri, B., and Tanzer, N. 2005. What people regard as unjust: Types and structures of everyday experiences of injustice, 1–17.

Nelissen, R. M. A., Leliveld, M. C., van Dijk, E., and Zeelenberg, M. 2011. "Fear and guilt in proposers: Using emotions to explain offers in ultimatum bargaining." *European Journal of Social Psychology*, *41*:78–85.

Nichols, Shaun. 2014. "Process Debunking and Ethics." *Ethics* 124:727–749.

Nozick, Robert. 1981. *Philosophical Explanation*. Cambridge, MA: Harvard University Press.

Ojha, Hardeo and Sah Brajbala. 1990. "Personality and socio-familial correlates of conservatism in Indian youth." *International Journal of Psychology* 25:295–304.

O'Mara, E. M., Jackson, L. E., Batson, C. D., and Gaertner, L. 2011. "Will moral outrage stand up?: Distinguishing among emotional reactions to a moral violation." *European Journal of Social Psychology* 41:173–179.

Polman, Evan, and Rachel L. Ruttan. 2012. "Effects of anger, guilt, and envy on moral hypocrisy." *Personality and Social Psychology Bulletin* 38:129–139.

Prinz, Jesse. 2011. "Against Empathy." *Southern Journal of Philosophy* 49:214–233.

Prinz, Jesse. 2014. "Forum: Against Empathy" *Boston Review*. Aug 26. http://www.bostonreview.net/forum/against-empathy/jesse-prinz-response-against-empathy-prinz

Pritchard, Duncan. 2005. *Epistemic Luck*. Oxford University Press.

Rakoczy, H., Warneken, F., and Tomasello, M. 2008. The sources of normativity: Young children's awareness of the normative structure of games. *Developmental Psychology*, 44:875–881.

Ray, John J. 1985. "What Old People Believe: Age, Sex, and Conservatism." *Political*

*Psychology* 6:525–528.

Richerson, P. J., and Henrich, J. 2012. Tribal Social Instincts and the Cultural Evolution of Institutions to Solve Collective Action Problems, Cliodynamics 3:38–80.

Riemann, Rainer, Claudia Grubich, Susanne Hempel, Susanne Mergl, Manfred Richter. 1993. "Personality and attitudes towards current political topics." *Personality and Individual Differences* 15:313–321.

Rosenthal-von der Pütten, A. M., N. C. Krämer, L. Hoffmann, S. Sobieraj, and S. C. Eimler. 2012. "An Experimental Study on Emotional Reactions Towards a Robot." *International Journal of Social Robotics* 5:17–34.

Rossano, F., Rakoczy, H., and Tomasello, M. 2011. "Young children's understanding of violations of property rights." *Cognition* 121:219–227.

Rottman, J., and Kelemen, D. 2012. "Aliens behaving badly: Children's acquisition of novel purity-based morals." *Cognition* 124:356–360.

Roush, Sherrilyn. 2005. *Tracking Truth: Knowledge, Evidence, and Science*. Oxford: Oxford University Press.

Roush, Sherrilyn. 2010. "The Value of Knowledge and Pursuit of Survival," *Metaphilosophy* 41:255–278.

Rozin, Paul, Laura Lowery, and Rhonda Ebert. 1994. "Varieties of disgust faces and the structure of disgust." *Journal of Personality and Social Psychology* 66: 870–881.

Ruse, Michael. 1986. *Taking Darwin Seriously*. New York: Blackwell.

Schafer, Karl. 2010. "Evolution and Normative Scepticism." *Australasian Journal of Philosophy* 88:471–488.

Schechter, Joshua. 2013. “Could Evolution Explain Our Reliability about Logic?” *Oxford Studies in Epistemology* 4:214–239.

Schmidt, M. F. H., Rakoczy, H., and Tomasello, M. 2012. Young children enforce social norms selectively depending on the violator’s group affiliation. *Cognition* 124:325–333.

Schnall, Simone, Jonathan Haidt, Gerald L. Clore, and Alexander H. Jordan. 2008. “Disgust as Embodied Moral Judgment.” *Personality and Social Psychology Bulletin* 34:1096–1109.

Sears, D.O. 1981. “Life Stage Effects Upon Attitude Change, Especially Among the Elderly.” In *Aging: Social Change*, ed. S. B. Kiesler, J. N. Morgan, and V. K. Oppenheimer, 183–204. New York: Academic Press.

Setiya, Kieran. 2013. *Knowing Right from Wrong*. Oxford: Oxford University Press.

Shafer-Landau, Russ. 2012. “Evolutionary debunking, moral realism, and moral knowledge.” *Journal of Ethics and Social Philosophy* 7:1–37.

Sher, George. 2001. “But I could be wrong.” *Social Philosophy and Policy* 18:64–78.

Singer, P. W. 2009. Wired for War: The Robotics Revolution and Conflict in the 21st Century. New York: Penguin Books.

Singer, Peter. 2005. “Ethics and Intuitions.” *The Journal of Ethics* 9:331–352.

Sinnott-Armstrong, Walter. 2006. *Moral Skepticisms*. Oxford: Oxford University Press.

Sinnott-Armstrong, Walter. 2008. “Framing Moral Intuitions.” In *Moral Psychology: Volume 2. The Cognitive Science of Morality: Intuition and Diversity*, ed. Walter Sinnott-Armstrong, 47–76. Cambridge, MA: The MIT Press.

Skarsaune, Knut Olav. 2011. “Darwin and moral realism: survival of the iffiest.” *Philosophical Studies*, 152:229–243.

Sosa, Ernest. 1999. “How to Defeat Opposition to Moore.” *Philosophical Perspectives* 13:141–54.

Street, Sharon. 2006. “A Darwinian dilemma for realist theories of value.” *Philosophical Studies* 127:109–166.

Street, Sharon. 2008. “Reply to Copp: Naturalism, Normativity, and the Varieties of Realism Worth Worrying About.” *Philosophical Issues* 18:207–228.

Stich, Stephen. 1990. *The Fragmentation of Reason*. Cambridge, MA: MIT Press.

Stouten, J., De Cremer, D., and van Dijk, E. 2005. “All is well that ends well, at least for proselfs: emotional reactions to equality violation as a function of social value orientation.” *European Journal of Social Psychology* 356:767–783.

Strohminger, Nina, Richard L. Lewis, and David E. Meyer. 2011. “Divergent effects of different positive emotions on moral judgment.” *Cognition* 119:295–300.

Swain, Stacey, Joshua Alexander, and Jonathan M. Weinberg. 2005. “The instability of philosophical intuitions: Running hot and cold on Truetemp.” *Philosophy and phenomenological research* 76:138–155.

Swinburne, Richard. 2004. *The Existence of God*. Oxford University Press.

Tersman, F. 2008. “The reliability of moral intuitions: A challenge from neuroscience.” *Australasian Journal of Philosophy* 86:389–405.

Thomson, N. R., and Jones, E. F. 2005. “Children's, Adolescents‘, and Young Adults’ Reward Allocations to Hypothetical Siblings and Fairness Judgments: Effects of Actor Gender, Character Type, and Allocation Pattern.” *The Journal of Psychology* 139:349–368.

Tiedens, Larissa Z., and Susan Linton. 2001. “Judgment under emotional certainty and uncertainty: the effects of specific emotions on information processing.” *Journal of personality and social psychology* 81: 973–978.

Trawalter, Sophie, K. M. Hoffman, and A. Waytz. 2012. “Racial Bias in Perceptions of Others’ Pain.” *PLoS ONE* 7: e48546.

Truett, K. R. 1993. “Age differences in conservatism.” *Personality and Individual Differences* 14:405–411.

Valdesolo, Piercarlo and David DeSteno. 2006. “Manipulations of emotional context shape moral judgment,” *Psychological Science* 17:476–77.

Valdesolo, Piercarlo, and David DeSteno. 2008. “The duality of virtue: Deconstructing the moral hypocrite.” *Journal of Experimental Social Psychology* 44:1334–1338.

van Inwagen, Peter. 1996. “It is wrong, everywhere, always, and for anyone, to believe anything upon insufficient evidence.” In *Faith, Freedom, and Rationality: Philosophy of Religion Today*, ed. Jeff Jordan and Daniel Howard-Snyder, 137–153 London: Rowman and Littlefield.

Vavova, Katia. ms. “Irrelevant Influences.” www.mtholyoke.edu/~evavova/files/vavova_ii.pdf. Accessed Sept. 12, 2014.

Vavova, Katia. "Debunking evolutionary debunking." In ed., Russ Shafer-Landau, *Oxford Studies in Metaphysics*, 9:76–101.

Vermunt, R., van Knippenberg, D., van Knippenberg, B., and Blaauw, E. 2001. “Self-esteem and outcome fairness: Differential importance of procedural and outcome considerations.” *Journal of Applied Psychology* 86:621–628.

Waddington, C. H. 1959. “Evolutionary systems: Animal and human.” *Nature* 183:1634–1638.

Warneken, Felix, et al. 2011. “Young children share the spoils after collaboration.” *Psychological Science* 22:267–273.

Waters, C. Kenneth. 2007. “Causes that make a difference.” *The Journal of Philosophy* 551–579.

Weinberg, Jonathan M. 2007. “How to challenge intuitions empirically without risking skepticism.” *Midwest Studies in Philosophy* 31:318–343.

Wielenberg, Erik J. 2010. “On the Evolutionary Debunking of Morality.” *Ethics* 120:441–464.

Wheatley, Thalia, and Jonathan Haidt. 2005. “Hypnotic disgust makes moral judgments more severe.” *Psychological Science* 16:780–784.

White, Roger. 2010. “You just believe that because . . . .” *Philosophical Perspectives* 24:573–615.

Zamzow, Jennifer L., and Shaun Nichols. 2009. “Variations in ethical intuitions.” *Philosophical Issues* 19:368–388.

Zhong, Chen-Bo, Brendan Strejcek, and Niro Sivanathan. 2010. “A clean self can render harsh moral judgment.” *Journal of Experimental Social Psychology* 46:859–862.

Zollman, K. J. S. 2008. “Explaining fairness in complex environments.” *Politics, Philosophy & Economics* 7:81–97.